395
7.9.70

SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

VOL. 22, PARTS 1 & 2 JANUARY 1960

CLASSIFICATION OF NATURAL AND PLANTATION TEAK (*TECTONA GRANDIS*) GROWN AT DIFFERENT LOCALITIES OF INDIA AND BURMA WITH RESPECT TO ITS PHYSICAL AND MECHANICAL PROPERTIES

By K. R. NAIR[1]

and

H. K. MUKERJI

Forest Research Institute, Dehra Dun, India

SUMMARY. The paper discusses the results of a multivariate analysis of the data on four important strength properties, namely, specific gravity, modulus of rupture, modulus of elasticity and maximum crushing stress for samples of teak consignments obtained from natural forests in eleven localities and from plantation forests in nine localities of the India-Burma region and tested for physical and mechanical strength properties at the Forest Research Institute, Dehra Dun. Following the procedure developed by C. R. Rao for forming group-constellations on the basis of Mahalanobis' D^2-statistic it was found that the localities of natural-grown teak could be classified into five groups and the localities of plantation-grown teak into three groups.

1. INTRODUCTION

A very large number of species of timber grows in the forests of India, which are of commercial importance. In the past, engineers and others believed in the use of only steel and reinforced concrete for all types of constructional work. They had no idea of the value of Indian timbers as good substitutes for steel for many purposes, until during the economic stress in the country, the resources of indigenous species of timber and their utility for a variety of purposes were exploited and established by the Forest Research Institute, Dehra Dun.

The Timber Mechanics Branch of the Forest Research Institute, which was established in 1921, has collected extensive data on the physical and mechanical properties of more than 250 indigenous species of timber. Most of these species are found in abundant and commercially exploitable quantities in the forests of the Indian Union. It should be mentioned here that in general five trees are used to estimate the average strength of a species from a particular locality or region.

---

[1] The author is now with the Central Statistical Organisation, Government of India, New Delhi.

The different types of tests of physical and mechanical strength that are carried out in the Timber Mechanics Branch of the Forest Research Institute and the strength properties which are derived from these tests are explained in the Appendix.

The Timber Mechanics Branch of the Forest Research Institute has made tentative classification of a large number of species of timber according to their suitability figures in a number of categories. For example, taking the property "strength as a beam," there are six categories, namely, (1) Extremely strong (2) Very strong (3) Strong (4) Moderately strong (5) Weak and (6) Very weak timbers (see Limaye 1957). According to these tentative classifications the usefulness of different species for a variety of purposes and as good substitutes to imported timber of foreign species are recommended.

The Timber Mechanics Branch of the FRI had collected in the course of several years strength data of 20 consignments of Teak (*Tectona grandis*) obtained from different localities of India and Burma. These consignments consisted of 11 natural-grown localities and 9 plantation-grown localities. Five trees in general were obtained from each locality for determining the strength properties. On the basis of his experience the Officer-in-Charge of the T. M. Branch made two tentative groups from 17 out of these twenty Teak-growing localities with regard to their physical and mechanical properties. One group represented a number of localities of South-West Burma, East and South India which is categorised as 'Strong' Teak. The other group of localities represented the Central Indian region including Madhya Pradesh and Bombay and categorised as 'Weaker' Teak (see Limaye 1946).

A few years ago, the entire data of 20 consignments of Teak were sent by the Officer-in-Charge, T. M. Branch to the first author for analysing the data in the Statistical Branch using modern statistical methods and to compare the results with his tentative grouping of Teak into two categories as mentioned above. The problem was therefore taken up by the authors and the application of standard methods for multivariate analysis by determining the discriminant functions and canonical variates adopted to classify the Teak consignments (localities) according to their physical and mechanical properties.

It will not be out of place to mention here that values of the large number of strength properties listed in the Appendix could not be determined individually from each specimen tested due to varying types of testing procedures some of which were destructive. It was therefore not possible to consider the individual specimens as the primary unit of observation of the strength properties for the analysis of data for this particular purpose. The only way was to consider for each strength property the mean value of all the specimens from a tree for that property as the primary unit of observation. Since there were not more than 5 trees in any consignment or locality only 4 strength properties which were considered very important by the Officer-in-Charge, T. M. Branch, from the point of view of their utility, have been selected in the present statistical investigation. These are (1) specific gravity (2) modulus of rupture

7. 9. 70

(3) modulus of elasticity in bending and (4) maximum crushing stress in compression parallel to grain test.

The analyses of data for the natural-grown and plantation-grown teak have been done separately in this paper.

2. VARIATION BETWEEN THE LOCALITIES WITH REGARD TO INDIVIDUAL PROPERTIES

The results of analysis of variance of a single property given in the earlier paper by the authors show that there is significant difference from locality to locality in the mean values of all the four properties of natural-grown teak and of two of the properties, namely, maximum crushing stress and modulus of elasticity of plantation-grown teak.

Tables 1 and 2 give the general mean values (mean of tree means) of the four properties, namely, specific gravity, modulus of rupture, modulus of elasticity and maximum crushing stress for 11 localities of natural-grown and 9 localities of plantation-grown teak respectively which have been used in the analysis. These mean values expressed as percentages of the general mean of all the localities in each type of growth (natural or plantation as the case may be) are also given in Tables 3 and 4 respectively. The large range of the mean values and the corresponding percentage figures in the two sets of tables suggest that there can be a number of strength-wise groups of the localities in each of the two types of growth, as against the popular classification into only two groups.

TABLE 1. VALUES OF GENERAL MEAN FOR DIFFERENT LOCALITIES (NATURAL-GROWN TEAK)

locality with symbol	no. of trees	specific gravity	modulus of rupture in units of 100 lbs/sq. in	modulus of elasticity in units of 10,000 lbs/sq. in	maximum crushing stress in units of 50 lbs/sq. in
North Kanara, Bombay (Typical Pole Forest) (N_1)	5	0.555	111.00	151.5	106.3
N. Kanara, Bombay (girdled) (N_2)	5	0.566	96.2	136.6	95.3
N. Kanara, Bombay (ungirdled) (N_3)	5	0.523	93.7	131.0	91.4
West Thana, Bombay (N_4)	5	0.576	109.1	144.5	108.6
Pyinmana, Burma (N_5)	5	0.613	143.2	190.9	150.5
Zigon, Burma (girdled) (N_6)	4	0.593	120.8	181.0	128.5
Zigon, Burma (ungirdled) (N_7)	5	0.588	114.6	164.4	113.8
Hoshangabad, M.P. (N_8)	5	0.554	103.3	140.2	104.2
South Chanda, M.P. (N_9)	5	0.527	90.9	120.7	89.6
Malabar, Madras. (N_{10})	5	0.615	107.7	156.9	109.4
S. Coimbatore, Madras (N_{11})	5	0.641	125.1	148.9	126.7
general mean		0.577	110.5	151.5	111.3

TABLE 2. VALUES OF GENERAL MEAN FOR DIFFERENT LOCALITIES (PLANTATION-GROWN TEAK)

locality with symbol	no. of trees	specific gravity	modulus of rupture in units in 100 lbs/sq. in	modulus of elasticity in units of 10,000 lbs/sq. in	maximum crushing stress in units of 50 lbs/sq. in
Chittagong Hill Tracts, Bengal (P_1)	5	0.556	119.7	171.4	129.1
Kurseong, Bengal (P_2)	5	0.538	115.4	165.3	121.0
Nilambur, Madras (Karumpuzha plantation) (P_3)	2	0.584	130.8	174.6	127.8
Nilambur, Madras (Panangode Plantation) (P_4)	3	0.547	114.9	144.5	109.5
Angul, Orissa (P_5)	5	0.608	112.2	155.9	114.9
C. Puri, Orissa (P_6)	5	0.583	111.4	150.5	110.8
S. Puri, Orissa (P_7)	5	0.533	99.1	127.7	96.6
Canal Beat, Hardwar, U.P. (P_8)	5	0.573	106.9	126.8	105.1
Pathri, Hardwar, U.P. (P_9)	5	0.569	113.3	156.8	108.2
general mean		0.565	112.5	151.5	112.8

TABLE 3. MEAN VALUES EXPRESSED AS PERCENTAGES OF THE GENERAL MEAN (NATURAL-GROWN TEAK)

locality with symbol	specific gravity	modulus of rupture	modulus of elasticity	maximum crushing stress
North Kanara, Bombay (typical pole forest) (N_1)	96	101	100	96
N. Kanara, Bombay (gridled) (N_2)	98	87	90	86
N. Kanara, Bombay (ungirdled) (N_3)	91	85	87	82
West Thana, Bombay (N_4)	100	99	96	98
Pyinmana, Burma (N_5)	106	130	126	136
Zigon, Burma (girdled) (N_6)	103	110	120	116
Zigon, Burma (ungirdled) (N_7)	102	104	109	103
Hoshangabad, M.P. (N_8)	96	94	93	94
South Chanda, M.P. (N_9)	91	82	80	81
Malabar, Madras (N_{10})	107	98	104	99
S. Coimbatore, Madras (N_{11})	111	113	99	114

TABLE 4. MEAN VALUES EXPRESSED AS PERCENTAGES OF GENERAL MEAN (PLANTATION-GROWN TEAK)

locality with symbol	specific gravity	modulus of rupture	modulus of elasticity	maximum crushing stress
Chittagong Hill Tracts, Bengal (P_1)	98	106	113	114
Kurseong, Bengal (P_2)	95	103	109	107
Nilambur, Madras (Karumpuzha plantation) (P_3)	103	116	115	113
Nilambur, Madras (Panangode plantation) (P_4)	97	102	95	97
Angul, Orissa (P_5)	108	100	103	102
C. Puri, Orissa (P_6)	103	99	99	98
S. Puri, Orissa (P_7)	94	88	84	86
Canal Beat, Hardwar (P_8)	101	95	84	93
Pathri, Hardwar (P_9)	101	101	104	96

3. ANALYSIS OF DISPERSION FOR THE TEST OF SIGNIFICANCE OF DIFFERENCES IN THE MEAN VALUES OF A NUMBER OF CORRELATED VARIABLES FOR SEVERAL LOCALITIES

In order to examine the above point, the first step taken was to apply the technique of analysis of dispersion into 'between' and 'within' localities and to test the mean values of different localities with Wilks' Λ Criterion. (See Rao, 1952, p.258).

The test of significance of differences in the mean value of several localities is based on the use of the statistic V where $V = -m \log_e \Lambda$ in which

$$\Lambda = \frac{|W|}{|S|} = \frac{\text{Dispersion matrix 'within localities'}}{\text{Dispersion matrix 'Total'}} \text{ and } m = n - \frac{p+q+1}{2}$$

$n = (n_1+n_2+\ldots+n_k)-1$ = Total number of observations less one.

p = number of properties taken.

k = number of localities and $q = k-1$.

Here, we are to use V as χ^2 with pq degrees of freedom.

Tables 5 and 6 give the analysis of dispersion for four characters, specific gravity (x_1), modulus of rupture (x_2), modulus of elasticity (x_3), and maximum crushing stress (x_4) for the data of 10 localities of natural-grown teak and 7 localities of plantation-grown teak, each consisting of 5 trees. It will be observed in both the tables that the value of the statistic V is highly significant (at 0.1% level of probability), which shows that the differences in the mean values with regard to the pooled effect of the above four properties between the different localities within the natural-grown and plantation-grown teaks are respectively highly significant.

TABLE 5. ANALYSIS OF DISPERSION (NATURAL-GROWN TEAK)

dispersion due to	d.f.	s. p. matrix									
		x_1^2	x_2^2	x_3^2	x_4^2	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
between localities	9	0.068034	11110.73	17165.22	14779.31	21.2506	23.2520	24.7753	12545.92	12722.33	14425.84
within localities	40	0.053971	5011.03	10674.78	4208.25	10.4491	12.0938	7.4778	5668.31	2880.94	3903.65
total	49	0.122005	16121.76	27840.00	18987.56	31.6997	35.3458	32.2531	18214.23	15603.27	18329.49

$V = 116.56^{***}$

TABLE 6. ANALYSIS OF DISPERSION (PLANTATION-GROWN TEAK)

dispersion due to	d.f.	s. p. matrix									
		x_1^2	x_2^2	x_3^2	x_4^2	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
between localities	6	0.020674	1301.29	9039.94	3410.06	1.1137	1.1058	0.6587	3179.34	1968.14	4989.61
within localities	28	0.036560	3178.78	5593.83	3068.67	2.7273	5.1068	2.7051	2927.80	2142.73	3093.53
total	34	0.057234	4480.07	14633.77	6478.73	3.8410	6.2126	3.3638	6107.14	4110.87	8083.14

$V = 63.69^{***}$

4. DIFFERENCES BETWEEN PAIRS OF LOCALITIES WITH REGARD TO MEAN VALUES OF THE FOUR PROPERTIES

Having known that the different localities within the natural-grown and plantation-grown teak differ significantly with regard to the four properties, values of D^2-(Square of Mahalanobis' generalised distance) between each pair of localities were determined and are given in Tables 7 and 8 for natural-grown and plantation-grown localities respectively.

The 'D^2' between two populations (localities) as estimated from the sample on the basis of p characters is given by

$$D_p^2 = \sum_1^p \sum_1^p (w^{ij}) \, [(\bar{x}_{i1}-\bar{x}_{i2}) \, (\bar{x}_{j1}-\bar{x}_{j2})]$$

where w^{ij} is the reciprocal matrix of w_{ij} the pooled dispersion matrix obtained from 'within localities' having 5 trees in each and $\bar{x}_{i1}$ and $\bar{x}_{i2}$ are the two sample means of the i-th property and $(i, j = 1, 2, \ldots p)$ (see Rao, 1952, p. 246).

For the test of significance of the mean values of p characters for any two populations (localities) out of k from which the dispersion matrix has been computed, the expression

$$\frac{n_1 n_2 (n_1+n_2+\ldots+n_k-k-p+1)}{p(n_1+n_2)(n_1+n_2+\ldots+n_k-k)} \cdot D_p^2$$

can be used as a variance ratio (F) with degrees of freedom p and $\{(\sum_1^k n_i)-k-p+1\}$.

In the case of problems of two samples only for which the dispersion matrix is calculated the above test reduces to using the statistic

$$\frac{n_1 n_2 (n_1+n_2-p-1)}{p(n_1+n_2)(n_1+n_2-2)} \cdot D_p^2$$

as a variance ratio with degrees of freedom p and (n_1+n_2-p-1). (See Rao, 1952, p.247).

When the properties under study are independent, i.e., they are mutually uncorrelated, the D^2 formula reduces to *simple sum of squares of the differences in the mean values of the various properties of the two populations.* It is therefore always desirable to transform the original variables, if they are mutually correlated by a set of mutually uncorrelated variables to simplify the computations.

This method suggested by Rao, 1952, p. 359 has been adopted in determining the various D^2 values which are given in Tables 7 and 8. The figures in a particular column of these tables indicate the square of the generalised distance between the locality given at the column head and all other localities corresponding to those figures in that column. The different number of stars against any figure show the levels of significance of difference between such pairs of localities. Figures without stars should be taken as *not significantly different from zero*. The symbols ($N_1, N_2, \ldots, N_{11}$) for natural-grown and ($P_1, P_2, \ldots, P_9$) for plantation-grown localities have been used in this paper.

From a scrutiny of the values of D^2 given in Table 7–natural grown–it will be seen that one of the Burma consignments (N_5) received from Pyinmana is the strongest of all and is significantly distant from the rest of the natural-grown teaks. Out of the two consignments from Madras, the one received from South Coimbatore (N_{11}) is significantly distant from the Malabar consignment (N_{10}) at 0.1% level of probability. It is also significantly distant from the rest of the localities. Similarly the consignment from Zigon, Burma (girdled) (N_6) is significantly distant from all the localities except that from Zigon, Burma (ungirdled) (N_7); on the other hand the latter consignment is very close to the Malabar consignment (N_{10}) from Madras. Four localities of Bombay (N_1 to N_4) and two localities of Madhya Pradesh (N_8 and N_9) appear to be very close to each other and can be grouped together. They have got lower strength than the rest of the localities under natural-grown.

Similarly it will be seen from Table 8 that the two Bengal teaks (P_1 and P_2) are very close to one another and one of the Nilambur teaks (Karumpuzha plantation) from Madras (P_3) is also close to them. The second Nilambur teak (Panangode plantation), Madras (P_4) is weaker and is significantly distant from Chittagong hill tracts, Bengal teak (P_1) and consequently it cannot be classed with the other three, rather it can come with those from Angul, Orissa (P_5), Central Puri, Orissa (P_6) and Pathri, Hardwar, U.P. (P_8), which are not significantly distant from each other.

5. Intra-and inter-cluster average D^2 and formation of tentative groups

The indications from the above paragraphs have been further explored by a study of Intra-and Inter-cluster average value of D^2 after the formation of tentative groups (clusters), such that the intra-cluster D^2 values are much less than the inter-cluster values, which gave a plausible rule for the formation of most appropriate grouping of different consignments (localities) within the *natural-grown* and *plantation-grown* teaks. A simple device following Rao (1952, p. 363) was to start with two closely associated groups having low D^2 values and find a third group which has the smallest average D^2 from the first two and so on. If at any stage the average D^2 of a group from those already listed appeared to be high, this group was taken out from the former cluster. If the change in average D^2 within a cluster due to inclusion of another group was found appreciable, then the newly added group has been taken outside the cluster.

Tables 9 and 10 give the average D^2 values 'within' and 'between' clusters which have been tentatively found out by the above device. No formal rules can however be laid down in the formation of these clusters, but the only criterion is that, on an average the D^2 values between any two localities within a group is much smaller than those belonging to two different groups. If a group consists of only one locality, this rule cannot be applied in such a case, as it is a separate group by itself.

TABLE 7. VALUES OF D^2 BASED ON FOUR PROPERTIES ARRANGED IN INCREASING ORDER OF MAGNITUDE WITH RESPECT TO EACH GROUP (NATURAL-GROWN TEAK)

N. Kanara Bombay (Typical Pole Forest) N_1	N. Kanara Bombay (girdled) N_2	N. Kanara Bombay (ungirdled) N_3	West Thana Bombay N_4	Pyinmana Burma N_5	Zigon, Burma (girdled) N_6	Zigon, Burma (ungirdled) N_7	Hoshangabad M.P. N_8	S. Chanda M.P. N_9	Malabar Madras N_{10}	S. Coimbatore Madras N_{11}
N_8 0.89	N_3 1.87	N_9 0.68	N_8 0.46	N_6 9.76**	N_7 2.95	N_1 1.77	N_4 0.46	N_3 0.68	N_7 2.60	N_4 6.63*
N_4 1.19	N_9 2.05	N_8 1.75	N_1 1.19	N_{11} 15.45***	N_1 7.14*	N_4 2.02	N_1 0.89	N_2 2.05	N_2 2.79	N_8 10.27***
N_7 1.77	N_8 2.20	N_2 1.87	N_7 2.02	N_7 16.96***	N_{10} 7.86**	N_8 2.40	N_3 1.75	N_8 2.24	N_4 4.07	N_7 10.83***
N_3 2.85	N_{10} 2.79	N_1 2.85	N_2 3.05	N_1 19.49***	N_4 7.99**	N_{10} 2.60	N_2 2.20	N_4 3.85	N_8 4.75*	N_1 11.62***
N_2 4.29	N_4 3.05	N_4 3.68	N_3 3.68	N_4 19.92***	N_8 8.64**	N_6 2.95	N_9 2.24	N_1 4.45	N_1 6.36*	N_{10} 12.05***
N_9 4.45	N_1 4.29	N_7 5.99*	N_9 3.85	N_8 23.08***	N_5 9.76**	N_2 4.48	N_7 2.40	N_7 8.27**	N_3 7.86**	N_6 15.13***
N_{10} 6.36*	N_7 4.48	N_{10} 7.86**	N_{10} 4.07	N_{10} 28.02***	N_2 13.45***	N_3 5.99*	N_{10} 4.75*	N_{10} 8.84**	N_6 7.86**	N_5 15.45***
N_6 7.14*	N_6 13.45***	N_6 15.01***	N_{11} 6.63*	N_3 35.69***	N_3 15.01***	N_9 8.27**	N_6 8.64**	N_{11} 17.62***	N_9 8.84**	N_2 15.77***
N_{11} 11.62***	N_{11} 15.77***	N_{11} 19.56***	N_6 7.99**	N_2 36.22***	N_{11} 15.13***	N_{11} 10.83***	N_{11} 10.27***	N_6 18.92***	N_{11} 12.05***	N_9 17.62***
N_5 19.49***	N_5 36.22**	N_5 35.69***	N_5 19.92***	N_9 38.98***	N_9 18.92***	N_5 16.96***	N_5 23.08***	N_5 38.98***	N_5 28.02***	N_3 19.56***

Note : * Significant at 5% – level of probability
** ,, ,, 1% ,, ,, ,,
*** ,, ,, .1% – ,, ,, ,,
and the rest are not significant at 5% – level of probability.

TABLE 8. VALUES OF D^2 BASED ON FOUR PROPERTIES ARRANGED IN INCREASING ORDER OF MAGNITUDE WITH RESPECT TO EACH GROUP (PLANTATION-GROWN TEAK)

Chittagong Hill Tracts Bengal P_1	Kurseong Bengal P_2	Nilambur Madras (Karumpuzha Plantation) P_3	Nilambur Madras (Panangode Plantation) P_4	Angul Orissa P_5	Central Puri Orissa P_6	South Puri Orissa P_7	Canal Beat, Hardwar U.P. P_8	Pathri, Hardwar U.P. P_9
P_2 0.77	P_1 0.77	P_1 3.01	P_6 1.76	P_6 0.64	P_5 0.64	P_4 2.29	P_7 3.52	P_6 1.31
P_3 3.01	P_3 3.62	P_2 3.62	P_7 2.29	P_9 2.64	P_9 1.31	P_6 3.39	P_4 3.73	P_5 2.64
P_6 5.53*	P_9 3.65	P_9 3.89	P_9 2.72	P_4 4.26	P_4 1.76	P_8 3.52	P_6 4.58	P_4 2.72
P_9 5.81*	P_6 4.59	P_6 4.13	P_8 3.73	P_3 4.57	P_7 3.39	P_9 4.80	P_5 6.26*	P_2 3.65
P_5 5.89*	P_4 4.92	P_5 4.57	P_5 4.26	P_1 5.89*	P_3 4.13	P_5 6.36*	P_9 9.48**	P_3 3.89
P_4 6.62*	P_5 5.98*	P_4 4.66	P_3 4.66	P_2 5.98*	P_8 4.58	P_2 8.50**	P_3 12.70*	P_7 4.80
P_7 11.71***	P_7 8.50**	P_7 12.10*	P_2 4.92	P_8 6.26*	P_2 4.59	P_1 11.71***	P_2 14.07***	P_1 5.81*
P_8 14.89***	P_8 14.07***	P_8 12.70*	P_1 6.62*	P_7 6.36*	P_1 5.53*	P_3 12.10*	P_1 14.89***	P_8 9.48**

Note : * Significant at 5% level of probability
** ,, ,, 1% ,, ,, ,,
*** ,, ,, .1% ,, ,, ,,
and the rest are not significant at 5% – level of probability.

TABLE 9. INTRA AND INTER-CLUSTER AVERAGE D^2
(NATURAL-GROWN TEAK)

group	localities	I	II	III	IV	V
I	N_5	—	15.4522	9.7595	22.4900	28.8978
II	N_{11}	15.4522	—	15.1311	11.4407	13.5785
III	N_6	9.7595	15.1311	—	5.4024	11.8575
IV	N_7 N_{10}	22.4900	11.4407	5.4024	2.6001	4.9657
V	$N_1N_2N_3N_4N_8N_9$	28.8978	13.5785	11.8575	4.9657	2.3672

TABLE 10. INTRA AND INTER-CLUSTER AVERAGE D^2
(PLANTATION-GROWN TEAK)

group	localities	I	II	III
I	$P_1P_2P_3$	2.4667	5.0212	12.3279
II	$P_4P_5P_6P_9$	5.0212	2.2220	5.1098
III	P_7P_8	12.3279	5.1098	3.5166

From the above analysis it has been found that the *natural*-grown teaks can be formed into five tentative groups and *plantation*-grown teaks into three tentative groups. The tentative groups arranged according to the descending order of their mean strength properties are as follows :

1. *Natural-grown teak.* Cluster 1 : Pyinmana, Burma (N_5). Cluster II : South Coimbatore, Madras (N_{11}). Cluster III : Zigon, Burma (girdled) (N_6). Cluster IV : (i) Malabar, Madras. (N_{10}); (ii) Zigon, Burma (ungirdled) (N_7). Cluster V : (i) North Kanara, Bombay (Typical pole forest) (N_1); (ii) North Kanara, Bombay (girdled) (N_2); (iii) North Kanara, Bombay (ungirdled) (N_3); (iv) West Thana, Bombay (N_4); (v) Hoshangabad, Madhya Pradesh (N_8); (vi) South Chanda, Madhya Pradesh (N_9).

2. *Plantation-grown teak.* Cluster I : (i) Chittagong Hill tracts, Bengal (P_1); (ii) Kurseong, Bengal (P_2); (iii) Nilambur, Madras (Karumpuzha plantation) (P_3). Cluster II : (i) Nilambur, Madras (Panangode plantation) (P_4); (ii) Angul, Orissa (P_5); (iii) Central Puri, Orissa (P_6); (iv) Pathri, Hardwar, U.P. (P_9). Cluster (III : (i) South Puri, Orissa (P_7); (ii) Canal Beat, Hardwar, U.P. (P_8).

The configuration of the clusters and their mutual relationships have also been graphically represented in Figures 1 and 2 for natural and plantation-grown teaks respectively. The square root of the average D^2 values represent the distance between and within groups, which have been calculated from the values in Tables 9 and 10. It is however not possible to represent all the distances to scale, hence they are only approximate.

CONFIGURATION OF CLUSTERS AND THEIR MUTUAL RELATIONSHIPS

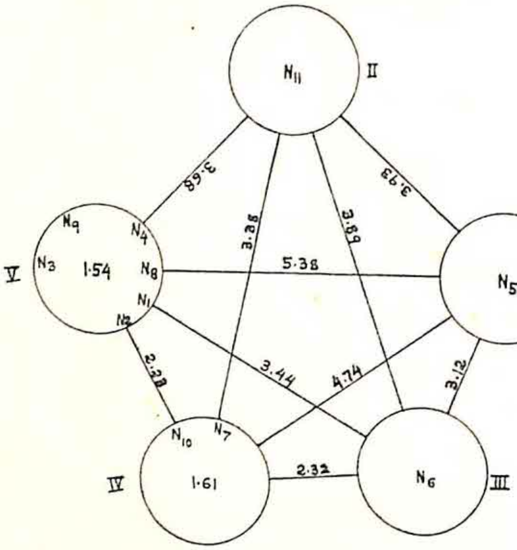


Figure 1.

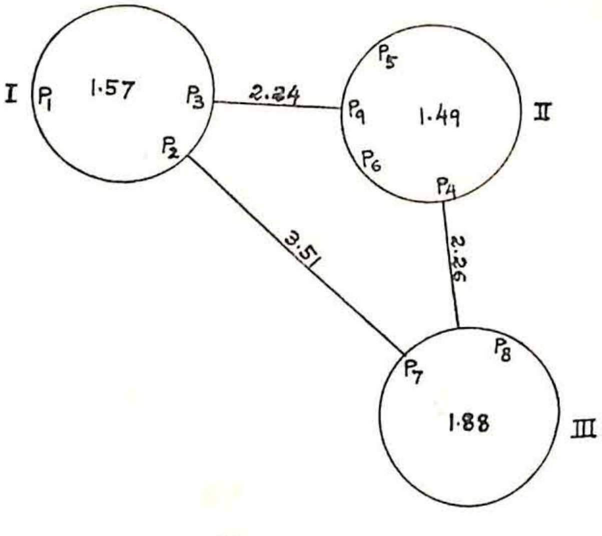


Figure 2.

6. Determination of group constellations by the use of canonical variates

This method deals with the replacement of the measurements of a number of mutually correlated characters (all of which contribute in some degree or other towards discrimination) by a relatively few measurements obtained as linear combinations of the large number of such measurements. The compounding coefficients are chosen in such a way that the new sets of variates can be arranged in a decreasing order of magnitude for discrimination. They are termed as canonical variates or canonical roots. This method of determining the group constellations is very useful when the number of characters is more than three and becomes much simpler if the different groups are characterised by two or three measurements which can easily be depicted on a chart, so that the underlying relationships between their forms are exhibited. It is closely related to but more flexible than the method discussed in the previous section for determining the group constellations (see Rao, 1952).

The main object of applying this method to the present problem of classifying the natural-grown and plantation-grown teaks (with regard to the variousp hysical and mechanical properties) growing in different localities of India and Burma is to examine how far the tentative classifications already arrived at by determining the intra-and inter-cluster average D^2 values as discussed in Section 5 agree with those determined by the use of this method.

In order to determine the D^2 between any two populations, the original set of the four correlated variables (mean of tree means for each locality), namely, specific gravity, modulus of rupture, modulus of elasticity and maximum crushing stress were transformed to an uncorrelated set following exactly the same steps as suggested by

Rao (1952). Tables 11 and 12 contain the set of uncorrelated transformed variables of the mean values for each locality under natural-grown and plantation-grown respectively. It will be seen that the values of the locality means of the four transformed characters are so adjusted that their sum for any character over all the localities (within natural-grown or plantation-grown) is zero, which needs no correction for calculating the raw sum of squares and products for "between localities."

The two Tables 13 and 14 representing the *Between product sum matrix* respectively for 'natural-grown' and 'plantation-grown' localities are given below. They have been obtained by using the mean values of the transformed variables given in the two Tables 11 and 12 respectively. These two tables are both (4×4) symmetric matrices, which have been represented by Matrix 'A' and Matrix 'B' respectively for natural-grown and plantation-grown localities.

TABLE 11. MEAN VALUES OF THE TRANSFORMED VARIABLES (NATURAL-GROWN TEAK)

locality	transformed variables			
	y_1	y_2	y_3	y_4
N_1	−0.617	+0.547	−0.014	−0.560
N_2	−0.314	−1.433	+0.082	−1.163
N_3	−1.499	−0.801	−0.112	−1.216
N_4	−0.028	−0.139	−0.547	−0.159
N_5	+0.981	+3.083	+0.300	+2.846
N_6	+0.430	+0.875	+1.793	+1.186
N_7	+0.292	+0.248	+0.828	−0.154
N_8	−0.645	−0.337	−0.306	−0.282
N_9	−1.388	−1.216	−0.846	−1.159
N_{10}	+1.036	−1.154	+0.777	−0.450
N_{11}	+1.752	+0.332	−1.955	+1.110

TABLE 12. MEAN VALUES OF THE TRANSFORMED VARIABLES (PLANTATION-GROWN TEAK)

locality	transformed variables			
	y_1	y_2	y_3	y_4
P_1	−0.275	+0.678	+1.461	+1.027
P_2	−0.785	+0.402	+1.389	+0.310
P^3	+0.518	+1.549	+0.515	+0.060
P_4	−0.530	+0.275	−0.780	−0.252
P_5	+1.198	−0.513	+0.117	+0.104
P_6	+0.490	−0.381	−0.149	−0.200
P_7	−0.926	−1.185	−0.916	−0.491
P_8	+0.207	−0.746	−2.080	+0.544
P_9	+0.093	−0.072	+0.443	−1.107

TABLE 13. BETWEEN PRODUCT SUM MATRIX USING MEAN VALUES OF THE TRANSFORMED VARIABLES

NATURAL-GROWN LOCALITIES : MATRIX 'A'

locality	transformed variables			
	y_1	y_2	y_3	y_4
y_1	10.445	6.082	0.225	9.064
y_2	6.082	16.380	2.326	14.522
y_3	0.225	2.326	9.544	1.536
y_4	9.064	14.522	1.536	15.558

TABLE 14. BETWEEN PRODUCT SUM MATRIX USING MEAN VALUES OF THE TRANSFORMED VARIABLES

PLANTATION-GROWN LOCALITIES : MATRIX 'B'

locality	transformed variables			
	y_1	y_2	y_3	y_4
y_1	3.825	0.290	−0.286	0.130
y_2	0.290	5.471	4.734	1.123
y_3	−0.286	4.734	10.335	1.028
y_4	0.130	1.123	1.028	3.031

According to Rao (1952) the canonical vectors $l_1, l_2, \ldots, l_4$ associated with a symmetric matrix are the same as those associated with a matrix formed of a suitable power of the original matrix. For example, if I is the vector associated with root λ of 'A' (original matrix for natural-grown localities) , then

$$I(A-\lambda I) = 0. \qquad \ldots \quad (1)$$

These equations give the maximum, minimum and intermediate values of λ.

Multiplying equation (1) by $(A+\lambda I)$ we get

$$I(A^2-\lambda^2 I) = 0. \qquad \ldots \quad (2)$$

Similarly multiplying equation (2) by $(A^2+\lambda^2 I)$ we get $I(A^4-\lambda^4 I) = 0$ and so on. Hence the results stated above hold good. It is also true that the convergence will be much quicker, if we start to solve the equations with a suitable power of the original matrix

The best procedure is to start with a trial vector (1, 1, ... 1) by multiplying each row of the matrix formed of a suitable power of the original matrix to obtain a derived vector which appears to be a better approximation than the trial one. (Rao, 1952).

In the present investigation as four characters have been used both in case of natural-grown and plantation-grown localities, there will thus be four latent roots (canonical roots) of λ. The sum of the four latent roots is the sum of the diagonal elements of matrices 'A' and 'B' for (natural-and plantation-grown teak). Hence $(\lambda_1+\lambda_2+\lambda_3+\lambda_4)$ is equal to 51.927 for natural-grown teak and 22.662 for plantation-grown teak. It has also been found that the sum of the first two canonical roots,

$\lambda_1+\lambda_2$ is equal to 45.05 for natural-grown teak, which alone accounts for a total variation of 86.8%. Similarly the first two canonical roots in case of plantation-grown teak is equal to 17.43 which accounts for a total variation of 76.9%. Hence these two roots contribute to a sufficiently high degree towards the discrimination of the various groups. The canonical or the standardised vectors corresponding to the first two canonical roots are given in Tables 15 and 16 for the natural-grown and plantation-grown localities.

TABLE 15. CANONICAL VECTORS CORRESPONDING TO THE FIRST TWO CANONICAL ROOTS

(NATURAL-GROWN TEAK)

canonical roots	transformed variables			
	y_1	y_2	y_3	y_4
λ_1	.3940	.6377	.0997	.6543
λ_2	−.3086	.1209	.9398	−.0835

TABLE 16. CANONICAL VECTORS CORRESPONDING TO THE FIRST TWO CANONICAL ROOTS

(PLANTATION-GROWN TEAK)

canonical roots	transformed variables			
	y_1	y_2	y_3	y_4
λ_1	−.0075	.5204	.8425	.1393
λ_2	.9189	.2631	−.1910	.2234

The mean values of the canonical variates for the different localities of natural-grown and plantation-grown teak calculated by using the mean values of the canonical vectors corresponding to $y_1, y_2, \ldots, y_4$ from Tables 15 and 16 are given in the Tables 17 and 18 respectively.

TABLE 17. MEAN VALUES OF THE CANONICAL VARIATES FOR 11 LOCALITIES OF NATURAL-GROWN TEAK

canonical variate	locality										
	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9	N_{10}	N_{11}
Z_1	−.262	−1.790	−1.908	−.258	4.245	1.683	.255	−.684	−2.165	−.545	1.433
Z_2	.290	.098	.362	−.509	.114	1.559	.731	−.106	−.417	.309	−2.43

TABLE 18. MEAN VALUES OF THE CANONICAL VARIATES FOR 9 LOCALITIES OF PLANTATION-GROWN TEAK

canonical variate	locality								
	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
Z_1	1.729	1.429	1.245	−.545	−.163	−.355	−1.450	−2.066	.181
Z_2	−.124	−.812	.799	−.322	.967	.334	−1.097	.513	−.265

The sum of squares of Z_1 and Z_2 correspond exactly with the values of λ_1 and λ_2 (the first two canonical roots determined earlier for both natural-grown and plantation-grown teak). They have thus provided a check on the calculations.

As already stated that about 87 percent and 77 percent of the total variations $(\lambda_1+\lambda_2+\lambda_3+\lambda_4)$ for natural-grown and plantation-grown teak are accounted for due to the first two canonical roots, they supply the best two linear functions

$$(l_{11}y_1+l_{21}y_2+l_{31}y_3+l_{41}y_4) = Z_1$$

$$(l_{12}y_1+l_{22}y_2+l_{32}y_3+l_{42}y_4) = Z_2$$

so that a two-dimensional representation with these two canonical variates as coordinates will give a fairly accurate picture of the configuration of the various groups. The third and fourth canonical roots contribute very little and hence may be neglected, although a three-or four-dimensional model showing the positions of the various localities will give the true picture.

The two-dimensional charts with the canonical variates as coordinates, showing the positions of the various localities under the natural-grown and plantation-grown teak are drawn in the Figures 3 and 4 respectively. From these two charts it is indicated that the 11 natural-grown teak localities as seen from the clustering can be classified into 5 broad groups and the 9 plantation-grown teak localities into 3 broad groups. The classifications are as follows :

1. *Natural-grown localities.* Group I : Pyinmana, Burma (N_5); Group II : South Coimbatore, Madras (N_{11}); Group III : Zigon Burma (girdled) (N_6); Group IV : (i) North Kanara, Bombay (Typical Pole Forest) (N_1); (ii) West Thana, Bombay (N_4); (iii) Zigon, Burma (ungirdled) (N_7); (iv) Hoshangabad, Madhya Pradesh (N_8); (v) Malabar, Madras (N_{10}). Group V : (i) North Kanara, Bombay (girdled) (N_2); (ii) North Kanara, Bombay (ungirdled) (N_3); (iii) South Chanda, Madhya Pradesh (N_9).

2. *Plantation-grown localities.* Group I : (i) Chittagong Hill tracts, Bengal (P_1); (ii) Kurseong, Bengal (P_2); (iii) Nilambur, Madras (Karumpuzha plantation) (P_3). Group II : (i) Nilambur, Madras (Panangode Plantation) (P_4); (ii) Angul, Orissa (P_5); (iii) Central Puri, Orissa (P_6); (iv) Pathri, Hardwar, U.P. (P_9). Group III : (i) South Puri, Orissa (P_7); (ii) Canal Beat, Hardwar, U.P. (P_8).

It may be mentioned here that there is some disagreement in the clustering of the different localities in case of natural-grown teak as obtained by the two different methods (see Figs. 1 and 3) namely, by the determination of the Inter-and Intra-cluster average D^2 and by the determination of the Canonical variates. By the latter method, three of the localities namely, North Kanara, Bombay (Typical pole forest) (N_1), West Thana, Bombay (N_4) and Hoshangabad, Madhya Pradesh (N_8) have come up to the higher group (IV) and classed with Malabar, Madras (N_{10}) and Zigon, Burma (ungirdled) (N_7). There is however close agreement between the two methods in the case of plantation grown teak (see Figs. 2 and 4).

GROUP CONSTELLATION OF LOCALITIES

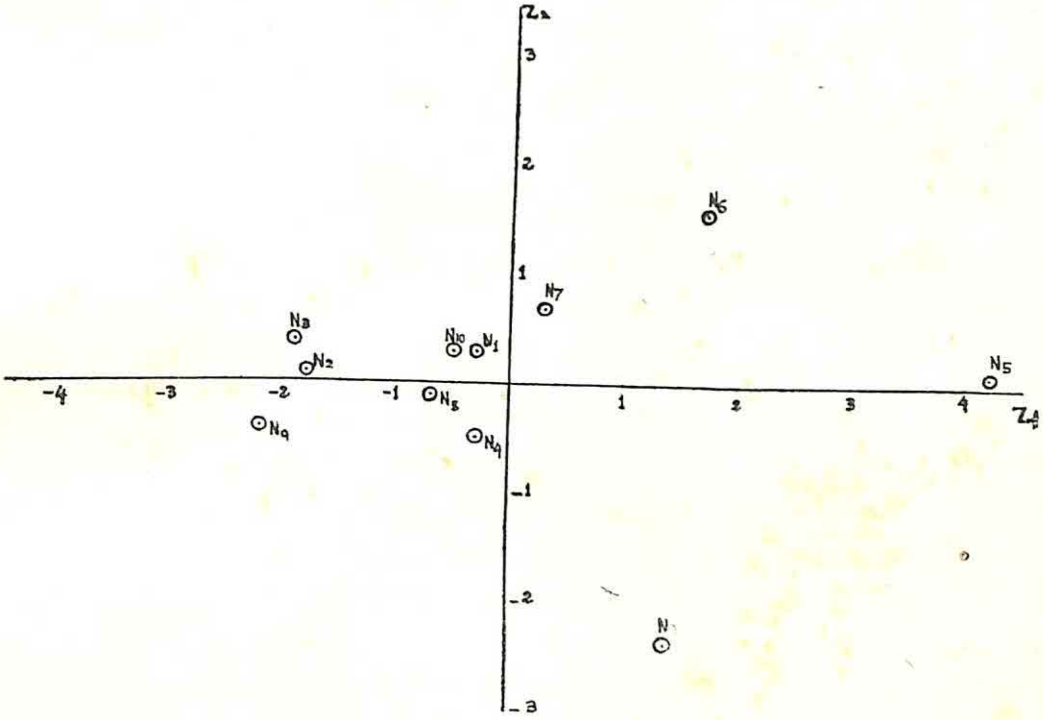


Figure 3. Natural-grown teak.

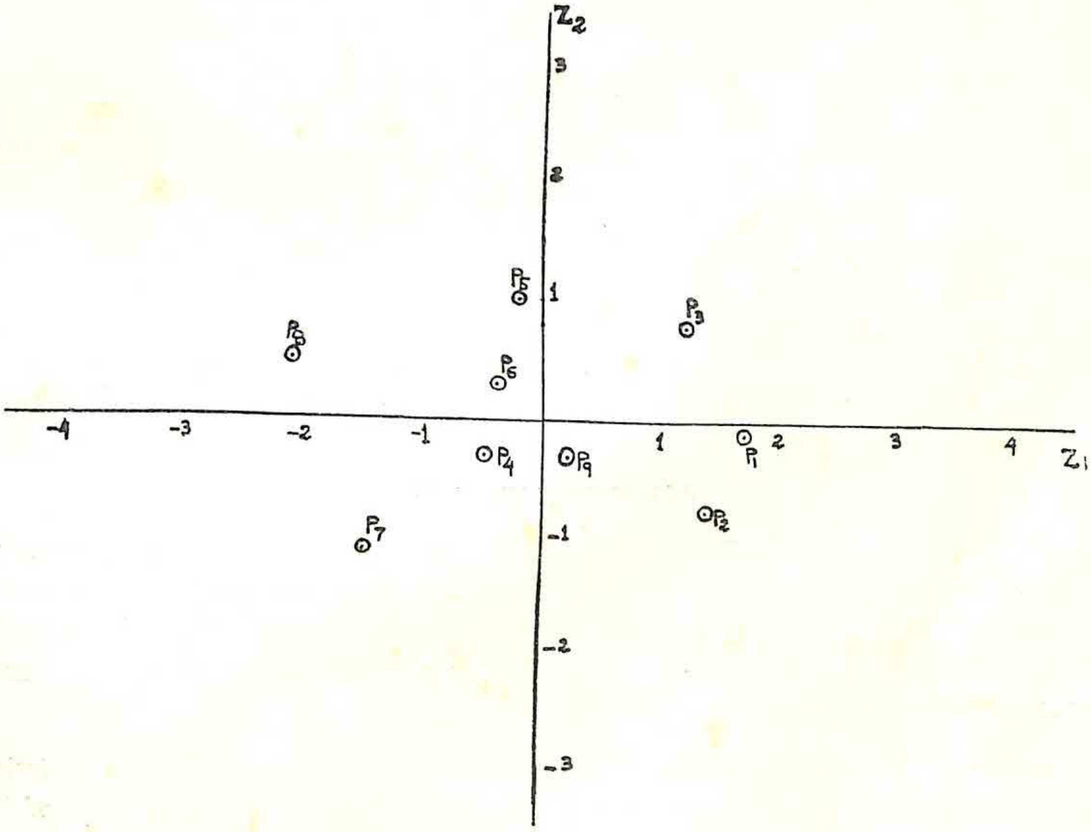


Figure 4. Plantation-grown teak.

7. SUMMARY OF RESULTS

(1) The mean values of four important characters, specific gravity, modulus of rupture, modulus of elasticity and maximum crushing stress for the various localities within the natural-grown and plantation-grown teaks showed a wide range of variation from locality to locality.

(2) The result of analysis of dispersion by Wilks' criterion showed that the pooled effect of the mean values of all the above four properties differed significantly from locality to locality both in natural-grown and plantation-grown teaks.

(3) Mahalanobis' D^2 statistic with respect to above four characters was tested for significant difference between different pairs of localities within the natural-grown and plantation-grown teak consignments. A large number of pairs of localities were found to differ significantly. These results were utilised to form tentative groups of the localities within each type according to their closeness or significant distance from each other.

(4) From the results of the Intra-cluster and Inter-cluster average D^2 values, the *natural-grown* teaks can be classified into five tentative groups and the *plantation-grown* into three tentative groups.

(5) From the results of the first two canonical roots it was found that about 87 percent of the total variation was accounted for by these two roots in case of natural-grown teak and about 77 percent in case of plantation-grown teak.

(6). From the two-dimensional representation of the different localities with the first two canonical variates as the axes of coordinates, it has been found that the localities of natural-grown teak can be classified into 5 broad groups and the localities of plantation-grown teak into 3 broad groups.

(7) Though the number of groups as determined by the two methods, namely from the study of Intra-and Inter-cluster average values of D^2 and the first two canonical roots are same, there has been a shifting of a few localities from one group to another in case of natural-grown teaks by the above two methods.

ACKNOWLEDGEMENTS

The data used in the present statistical study were accumulated in the Timber Mechanics Branch of the Forest Research Institute over a long period of years and were sent to the Statistical Branch for a detailed study at the request of the senior author by the late shri V. D. Limaye former Officer-in-Charge of the T. M. Branch to whom our thanks are due.

The authors also sincerely thank the entire computing staff of the Statistical Branch for their wholehearted assistance from time to time in the computation of the huge bulk of data which took pretty long time to complete.

Appendix

TYPES OF TESTS

(1) *Static bending test* carried out on small clear specimens of size 30″ × 2″ × 2″ supported at two ends and loaded centrally.

The strength properties derived from this test are: (i) Fibre stress at elastic limit (ii) Modulus of rupture (iii) Modulus of elasticity (iv) Work to elastic limit (v) Work to maximum load and (vi) total work.

(2) *Impact bending test* carried out on small clear specimens of size 30″ × 2″ × 2″ supported at two ends and a hammer of 50 lb. weight is allowed to fall freely from successively increasing heights at the middle of the specimen till the specimen breaks. The strength properties derived from this test are: (i) Fibre stress at elastic limit (ii) Modulus of elasticity (iii) Work to elastic limit and (iv) Maximum height of drop.

(3) *Compression parallel to grain test* carried out on small clear specimens of size 8″ × 2″ × 2″ axially loaded along grain. The strength properties derived from this test are (i) Compressive stress at elastic limit (ii) Maximum crushing stress and (iii) Modulus of elasticity.

(4) *Compression perpendicular to grain test* carried out on small clear specimens of size 6″ × 2″ × 2″ by applying load through a 2″ wide thick steel plate placed on the middle of the radial surface of the specimen. Only fibre stress at elastic limit is derived from this test.

(5) *Hardness test* carried out on small clear specimens of size 6″ × 2″ × 2″ by ascertaining the load required to imbed a steel ball of 0.444″ diameter upto half its diameter on the radial, tangential and end surfaces of each specimen seprately.

(6) *Shear test* carried out on small clear specimens of size 2″ × 2″ × 2″ having a projecting lip which is sheared off from the rest of the specimen by the application of a load on it. This test is carried out one on radial and another on tangential surface of two different specimens which are paired. Only shearing stress is determined from this test.

(7) *Tension perpendicular to grain test* carried out on small clear specimens of size 2″ × 2″ × 2″ in which a tensile load is applied at right angles to the grain by clutching the specimens which are notched. This test is carried out one on radial and another on tangential surface of two different specimens which are paired. Only tensile stress perpendicular to grain is determined from this test.

Besides these mechanical strength properties, a few physical properties namely, specific gravity, rings per inch and summer-wood % are also determined. The last two properties are determined from the specimens of only the compression parallel to grain tests stated in (3) and the specified gravity determination is done from almost all the specimens under the tests stated in (1) to (5). In addition to these tests a few other tests are also carried out, viz. shrinkage tests (volumetric, radial and tangential) and torsion tests.

Structural bending tests on big-sized rectangular beams having a span of about 15′ and varying cross sections upto a maximum of 6″ × 12″ and compression parallel to grain and perpendicular to grain tests on bigger dimensions than those in small clear specimens are also carried out for comparing their results with those obtained from small clear specimens. These strength figures are used along with the former ones for determining safety factors and safe working stress figures which are of immense help to the engineers for wooden constructional purposes (see Limaye 1942, 1946).

The large number of properties obtained from the above tests on small clear specimens are further made use of in determining the suitability of a species of timber for different purposes (see, Limaye, 1933 & 1944). A number of strength properties are combined after applying some weighting and adjusting factors to them according to their importance for the determination of their suitability for a particular purpose. There are seven types of suitability figures which can be made use of depending on the purpose for which a timber is to be used. They are as detailed below :

(1) *Strength as a beam* for which modulus of rupture, fibre stress at elastic limit in static bending test and fibre stress at elastic limit in impact bending test are used,

(2) *Stiffness as a beam* for which modulus of elasticity from both static bending and impact bending tests are used.

(3) *Suitability as a post or strut* for which maximum crushing stress and crushing stress at elastic limit from compression parallel to grain tests and modulus of elasticity in static bending tests are made use of.

(4) *Shock resisting ability* for which work to maximum load and total work in static bending tests and maximum height of drop in impact bending tests are used. This property also gives an idea of the toughness of the timber.

(5) *Retention of shape* for which volumetric shrinkage from green to oven dry, radial and tangential shrinkage from green to oven dry and ratio of tangential to radial shrinkage are used. The reciprocal of the figure obtained after applying the adjusting and weighting factors multiplied by 100 is used for determining retention of shape.

(6) *Shear* for which radial and tangential shearing stresses are made use of for determining the shear property.

(7) *Hardness* for which fibre stress at elastic limit in compression perpendicular to grain, radial hardness, tangential hardness and end hardness are made use of in determining the hardness figures.

The final suitability figures are the weighted averages obtained by giving a weightage of 2 to the figures obtained as above from the green strength properties and a weightage of 1 to those obtained from the air dry strength properties. These suitability figures are then expressed as a percentage of similar figures obtained for *Standard Teak*, which is at present taken from the average of two consignments of Zigon (Burma teak) and one consignment of Malabar (Madras teak). These figures are termed as "comparative suitability figures". The average values of each property for a particular species based on mean of tree means are used as the basic figures for determining these suitability figures.

It has been found from experience that for a particular strength property or even in more than one strength property or physical property, a number of species of timbers show similarity. On the other hand the same timbers show a large variation from species to species with regard to a number of the remaining physical and mechanical properties. It has even been found that a good deal of variation exists within the same species tested from different localities.

References

Limaye, V. D. (1933): The physical and mechanical properties of woods grown in India. Third interim Report on Project 1, *I.F.R.* XVIII, X.

——— (1942): Safe working stresses for some important Indian timbers. *I. F. Leaflet* No. 13.

——— (1944): Suitability and selection of timbers for different uses. Parts I and II, *I.F.R.* **3**, No. 5.

——— (1946): Safe working stresses for Indian timbers. *I.F.R.* **4**, No. 1.

——— (1957): Grouping of Indian timbers and their properties, uses and suitability. *I.F.R.* **1** No. 2.

Nair, K. R. and Mukerji, H. K. (1957): A statistical study of the variability of physical and mechanical properties of *Tectona grandis* (teak) grown at different localities of India and Burma and the effects of the variability on the choice of sampling plan. Presented in the Section of Statistics, Indian Science Congress, Calcutta (see Part IV Late Abstracts in the proceedings of the 44th Session, Calcutta, 1957, p.(3)). Full paper being published in Indian Forest Record, Vol. 1, No. 1 (Statistical).

Rao, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York.

Paper received : March, 1959.

THE NATIONAL SAMPLE SURVEY

SEVENTH ROUND : OCTOBER 1953—MARCH 1954

NUMBER 13

REPORT ON HOUSEHOLD TRANSPORT OPERATIONS

CONTENTS

		PAGE
FOREWORD	: by P. C. Mahalanobis	23
CHAPTER ONE	: Scope, coverage, design and summary of results ...	25
CHAPTER TWO	: Employment	31
CHAPTER THREE	: Fixed capital	38
CHAPTER FOUR	: Costs and earnings	43
CHAPTER FIVE	: Description of tables	49
APPENDIX I	: Tables (operating households)	55
APPENDIX II	: Tables (all households in transport operations) ...	67
APPENDIX III	: Comparison of results of sub-samples	81
APPENDIX IV	: Facsimile of the Schedule of Investigation ...	83

INDEX TO TABLES IN THE TEXT

CHAPTER TWO

TABLE (2.1)	: Estimated number and percentage of persons engaged in household transport enterprise in rural and urban areas : October 1953—March 1954	31
TABLE (2.2)	: Estimated number of households engaged in household transport enterprise in rural and urban areas : October 1953—March 1954	32
TABLE (2.3)	: Estimated number of persons engaged in household transport enterprise per working household: October 1953—March 1954	32
TABLE (2.4)	: Household and hired labour in terms of man-days per working day in rural and urban areas : October 1953—March 1954	32
TABLE (2.5)	: Monthly charges for hired labour in rural and urban areas : October 1953—March 1954	33

J-395

PAGE

Table (2.6) : Number of working persons by types of transport in rural areas : October 1953—March 1954 34

Table (2.7) : Number of working persons by types of transport in urban areas : October 1953—March 1954 34

Table (2.8) : Number of persons engaged in transport enterprise as principal means of livelihood : Census 1951 35

Table (2.9) : Number of persons engaged in transport enterprise as secondary means of livelihood : Census 1951 36

Table (2.10) : A comparison of number of persons other than employees engaged in household transport enterprise as obtained from 1951 Census and NSS seventh round .. 37

CHAPTER THREE

Table (3.1) : Percentage of sample households which inherited land and buildings used in transport enterprise : October 1953—March 1954 38

Table (3.2) : Values of land and buildings per household used in transport enterprise in rural areas : October 1953—March 1954 39

Table (3.3) : Values of land and buildings per household used in transport enterprise in urban areas : October 1953—March 1954 39

Table (3.4) : Values of animals : vehicles and tools per household used in transport enterprise in rural areas, October 1953—March 1954 40

Table (3.5) : Values of animals, vehicles and tools per household used in transport enterprise in urban areas : October 1953—March 1954 41

Table (3.6) : Breakdowns of fixed capital in values and percentages for different types of transport in rural areas : October 1953—March 1954 41

Table (3.7) : Breakdowns of fixed capital in values and percentages for different types of transport in urban areas : October 1953—March 1954 42

CHAPTER FOUR

Table (4.1) : Breakdowns of monthly costs per operating household in transport enterprise in rural areas : October 1953—March 1954 43

Table (4.2) : Gross earnings per month per household by goods and passenger traffic in rural areas : October 1953—March 1954 45

Table (4.3) : Monthly earnings and costs per household in transport enterprise in rural areas : October 1953—March 1954 46

Table (4.4) : Breakdowns of monthly costs per operating household in transport enterprise in urban areas : October 1953—March 1954 46

Table. (4.5) : Gross earnings per month per household by goods and passenger traffic in urban areas : October 1953—March 1954 47

Table (4.6) : Monthly earnings and costs per household in transport enterprise in urban areas : October 1953—March 1954 48

THE NATIONAL SAMPLE SURVEY

SEVENTH ROUND : OCTOBER 1953—MARCH 1954

NUMBER 13

REPORT ON HOUSEHOLD TRANSPORT OPERATIONS

FOREWORD

0.1. The National Sample Survey was started in 1950 with the object primarily of filling gaps in statistical data, and special emphasis has continued to be given in successive rounds of surveys to the collection of information relating to the household sector of the national economy. Productive enterprises, which were owned and operated singly or jointly with other members of the same household or other households, were classified, as a matter of convenience, into agriculture and animal husbandry, manufacture and handicrafts, transport, trade, and professions.

0.2. Information on transport services rendered by households has been collected from the very beginning, but the size of the sample was small up to the Sixth Round. In the Seventh Round (October 1953—March 1954), a relatively large number of households engaged on transport services was investigated and findings based on this survey have been given in this report.

0.3. The data collected have reference only to that part of the transport services which is sold for cash or kind. This is a severe limitation because it is possible that a large volume of transport services in rural areas is used by the owners themselves which are excluded from the present survey.

0.4. The total number of households engaged in transport services was comparatively small, only 1.46 million, of which 1.06 million (73 per cent) were in rural areas. Owing to the seasonal nature of the activity only about half of the households were actually working during the reference period of the survey. In all, 1.1 million persons were engaged in transport services which is 0.78 per cent of the total working population. The monthly net earning per working household was Rs. 20.46 (46 per cent of Rs. 44.10, the average gross earning per household) in rural areas; and Rs. 55.16 (or 62 per cent of the gross earning of Rs. 89.19) in urban areas.

0.5. It will be seen that the household transport sector is a small part of the national economy. The present data are, however, not complete as they refer only to that part of the transport services which is rendered for compensation received in cash or in kind. The transport services utilised for the owners themselves, which have been excluded from the present survey, may, however, be quite appreciable in rural areas. Although this report thus gives only a partial view of the sector, being the first study of its kind, it would have interest and value.

0.6. The report was prepared under the direction of Satyabrata Sen by K. G. Chandrasekharan Nair, Bijankumar Datta Roy, S. Ramanatha Iyer and others in the Household Enterprise Section in the Department of National Sample Survey of the Indian Statistical Institute (ISI). The final form is naturally the result of co-operative effort of men in both Statistical and Field Wings of the National Sample Survey (NSS). Specific mention may, however, be made of Debabrata Lahiri and Nimai Charan Ghosh of ISI for design and planning of the survey; Haricharan Ghosh, Chief Director, NSS, for field work, and Birendra Mohan Das Gupta of ISI for editing.

1 November 1958 P. C. Mahalanobis

THE NATIONAL SAMPLE SURVEY

SEVENTH ROUND: OCTOBER 1953—MARCH 1954

NUMBER 13

REPORT ON HOUSEHOLD TRANSPORT OPERATIONS

*This Report on Household Transport Operations was prepared by the Indian Statistical Institute and is being published in the form in which it was submitted to the Government of India. The views contained in this report are not necessarily those of the Government of India.**

CHAPTER ONE

SCOPE, COVERAGE, DESIGN AND SUMMARY OF RESULTS

INTRODUCTION

1.1. Transport service as a vital constituent in the productive activity, no doubt, plays a very significant role in the present-day economic system. The total facilities in India, however, consist not only of mechanised means of transport by land, sea and air but also of bullock carts and similar other forms of transport operated by individual households. Some details are available in respect of certain sections of large-scale and mechanised transports such as the railways, but information relating to small-scale and unorganised means of transport is almost absent. Animal and animal-drawn transport and transport by small wooden boats, though antiquated, are still abundant in India, and so also is carrying of goods by human beings. The bullock carts, horses, ponies and camels serve the remote parts of this big country, not accessible to any other form of transport. In urban areas also the non-mechanised forms of transport are used to a great extent.

1.2. Data on small-scale transport operations were collected in the National Sample Survey by completing questionnaires from a sample of households engaged in transport operations. This report is based on particulars collected in the seventh round of the survey. In the four previous rounds, viz., third, fourth, fifth and sixth, the sample households for survey of different kinds of enterprises were selected from a common list of households in the sample villages. The proportions of households following different kinds of occupations as they existed in the population were reflected in the sample. Since the proportion of households engaged in transport operations was very small, the number included in the sample was also very small. Consequently, in these four rounds, the effective sample sizes were very small and it was not worthwhile to take up detailed tabulation of these data. In the seventh round, to obviate this difficulty, separate lists of households were prepared for different occupations and a large number of samples were selected from households which were offering transport services. In all, 4394 sample households engaged in transport operations were interviewed and this larger sample size accounts for the preference for tabulation of the seventh round data to those of the previous rounds.

---

*The draft report (No. **D. 17**) was submitted to the Government of India in December 1956.

1.3. The seventh round covered a period of six months from October 1953 to March 1954. Particulars were collected from 2742 households chosen from 959 sample villages and 1652 households chosen from 443 urban sample blocks covering the country. One sample village and a sample block had, however, to be treated as casualties. The manner of selection of villages, blocks and households, and concepts and definitions have been described below.

Scope of the survey

1.4. The estimates presented in this report are based on data collected in respect of all transport enterprises undertaken either by individual households or jointly by more than one household. By transport enterprise is meant the offering of transport service to others as a business proposition. The common examples are the carrying of passengers or goods on payment of hire charges by means of a vehicle or by a person or an animal. Thus, porters selling transport services have also been brought under the scope of this survey. Those households which rented out their transport vehicles to others and did not operate them directly with household or hired labour have been left out since the very act of hiring out their means of transport makes them rent-receivers. Again, those households have been excluded which used transport equipment exclusively for some other enterprises of their own and did not hire the services out. Government agencies, joint stock companies, co-operatives and similar other non-household enterprises engaged in transport operations were also outside the scope of the survey.

1.5. It may be mentioned here that if any sample household carried on joint activities of transport and repairing services, the data regarding transport enterprise and repairing services were collected separately and the common costs between the two were allocated in proportion to their gross earnings.

1.6. All the particulars regarding land, buildings, draught and pack animals, transport vehicles, consumption of fuel, lubricants and cattle feed and gross earnings have been collected and tabulated. Information collected relates, in general, to the reference period of a month preceding the date of survey. In the case of consumption of cattle feed, fuel, lubricants and other materials data for the reference period of a week preceding the date of survey also have been collected. The tables presented in this report, however, give only monthly estimates. No attempt has been made to estimate annual values lest the seasonal factors should vitiate the accuracy of the results.

1.7. All households who derived the whole or a part of their income by rendering transport service were included in the frame for selection of sample households irrespective of whether they were actually engaged in the enterprise during the month prior to the date of investigation or not. Thus, the households who were not rendering transport service in the reference period but regarded themselves to be usually in the occupation were also selected for survey in view of the seasonal nature of the enterprise. Investigations for the seventh round were undertaken in October 1953 and continued up to March 1954. During this period agricultural operations in respect of kharif crops were at their height and hence it is likely that activities in respect of transport enterprise remained suspended in all those households where agriculture provided the chief means of livelihood and transport service occupied a subsidiary position. In regard to such households, information on employment and gross earnings could not be had as the reference period to which the data were to pertain was a month preceding the date of investigation. However, the data on fixed capital and cost of upkeep of animals maintained for transport enterprise were available.

1.8. Because of this seasonal variation two sets of estimates have been prepared. One set relates to all households offering transport services regardless of their stoppage of activities at certain periods of the year and the various particulars relating to employment, expenses and income have been spread over all these households. These estimates will be useful later to study seasonal fluctuations when data relating to different points of time will be collected (Appendix II). The other set relates to only those households which were offering transport services in the reference period of this survey and excludes all those which were not operating in this period. The particulars of employment, expenses and income of these operating households have been shown separately (Appendix I). To make the position clear, the tables have been given in two separate sets in Appendices I and II.

1.9. A point about the data relating to fixed capital should be noted. In the case of land and buildings, the 'cost of acquisition' and not the value on the date of survey was collected. Similarly, for transport vehicles, tools, draught and pack animals etc., the value at first use has been collected. The 'cost of acquisition' or 'the value at first use' will not, however, admit of any comparison on account of the changes in price at different times when similar assets may have been acquired by different households. This procedure, in spite of its limitations, was adopted due to some practical difficulty of determining their values on the date of survey. However, it may be mentioned here that this method has been adopted in other major surveys like the Census of Manufacturing Industries, also. In cases where land and buildings etc. were inherited by the present operator, the cost of acquisition could not be ascertained. Hence, the data relating to fixed capital items given in this report have reference to the cost of acquisition of non-inherited part only.

COVERAGE

1.10. As in the four earlier rounds, the geographical coverage of the seventh round of the survey was the whole of rural and urban areas of the Indian Union excluding Jammu and Kashmir, and Andaman and Nicobar Islands. Altogether more than 40,000 sample households were visited for the different enquiries undertaken in this round. The number of households contacted for particulars of transport enterprise were 4394 of which 2742 were from rural areas and 1652 from urban areas.

DESIGN OF THE SURVEY

1.11. *Rural areas :* The sampling was done in three stages. In the first stage, sample tehsils were selected after suitably stratifying all the tehsils into 240 strata. In the second stage, sample villages were selected from each selected tehsil and in the third stage, sample households were selected from each of the sample villages.

1.12. India was first divided into 'Natural Divisions' as defined by the Registrar General, Ministry of Home Affairs, Government of India. The total consumer expenditure for each Natural Division was then calculated on the basis of data collected in the first round of National Sample Survey and the allocation of the number of tehsils to be sampled from the different Natural Divisions was proportional to these consumer expenditures. Stratification of tehsils within each Natural Division was done on the basis of density of population subject to the condition that total consumer expenditure of each stratum within the Natural Division would be more or less equal. The number of strata so formed in each Natural Division was such that two sample tehsils were to be selected from each stratum. Thus, the total consumer expenditures of all strata of all Natural Divisions were approximately equal. The number of strata, so constituted, was 240.

1.13. From each stratum, two sample tehsils were selected with probability proportional to area or population of the tehsils and with replacement. Within each selected tehsil two sample villages were then selected again with probability proportional to area or population of the villages and with replacement. Thus, in all, 960 sample villages were selected for survey.

1.14. In each sample village all the households were listed with indications for broad industrial grouping by the productive enterprise of the household. The list of all households engaged in transport enterprise was then prepared and from this 10 sample households were selected for survey on details regarding transport enterprise. In those villages where the number of households engaged in transport enterprise was less than 10, all the households were selected. The total number of households so selected and surveyed was 2742.

1.15. *Urban areas :* As in rural area the sampling in urban area was also done in three stages. In the first stage, sample towns were selected after proper stratification of all the towns. In the second stage, sample blocks were selected from each first stage unit and in the third stage, sample households were selected from each sample block.

1.16. All the towns in India excluding the four big cities of Bombay, Calcutta, Delhi and Madras were divided into four strata based on the population of the towns and from each stratum towns were selected separately from four regions constituting the whole of India. These four regions were built up from the six population zones adopted by the Registrar General. Towns were selected with probability proportional to the population of the towns and with replacement. The number of towns so selected was 53 of which 14 were from towns with population 100,000 and over; 9 from towns with population between 50,000 and 100,000; 14 from towns with population between 15,000 and 50,000; and 16 from towns with population below 15,000. Besides these 53 towns, the four big cities of Bombay, Calcutta, Delhi and Madras with population over 1,000,000 were also included in the survey making in all 57 sample towns and cities.

1.17. From each of these 57 urban units, required number of blocks, depending upon the size class, were selected with probability proportional to the total number of households in the blocks and with replacement. The total number of blocks so selected were 444. Procedure of selection of households for particulars of transport enterprise in each sample block was exactly similar to that in the villages. In all, 1652 households were surveyed.

1.18. All-India estimates for rural sector were obtained by summing up all the stratum estimates, which were got first by getting the estimates of totals of all households engaged in the transport enterprise in the sample villages and then by combining them using suitable weights based on the respective probabilities of selection.

1.19. The procedure of estimation for urban areas is exactly similar to that for rural areas. First, the estimates of totals of all households in transport enterprise in the sampled blocks were obtained and they were coalesced together to give stratum estimates making use of proper weights depending on the corresponding probabilities of selection. All-India estimates were got by adding all the stratum estimates.

SUMMARY OF MAIN RESULTS

1.20. *Households :* There were 1.46 million households in India engaged in household transport enterprise as estimated from the seventh round of the National Survey (October 1953–March 1954). 1.06 million of these households (73 per cent) were in rural areas and the remaining 0.40 million (27 per cent) in urban areas.

1.21. Due to seasonal nature of the activity only 51 per cent of the households in rural areas (0.55 million) and 92 per cent in urban areas (0.37 million) were actually working in transport enterprise during the reference period of the survey.

1.22. *Type of transport :* In rural areas out of the total number of 1.06 million households 0.91 (86 per cent) million were operating animal-drawn transport or animal transport and all other types of transport were of minor importance.

1.23. In urban areas animal-drawn and animal transport accounted for the activity of 0.12 million households (30 per cent) out of the total number of 0.40 million households, man-driven transport like push cart, rickshaw for 0.14 million households (35 per cent) and transport coolie and porter activities for 0.12 million households. (30 per cent).

1.24. *Employment :* In all, 1.11 million persons were engaged in household transport enterprise which is only 0.78 per cent of the total working population. Figures separately for rural and urban areas are 0.67 million and 0.44 million respectively.

1.25. The average employment per working household for the country as a whole was 1.22 and the rural and urban averages were 1.23 and 1.21 respectively.

1.26. *Fixed capital :* Cost of fixed capital employed in household transport enterprise excluding land and buildings was Rs. 436 per household in rural areas of which the larger part viz., Rs. 275 was accounted for by cost of animals and only the remaining Rs. 161 was cost of vehicles and tools. But, in urban areas the per household cost of animals was only Rs. 78 while that of vehicles and tools was higher at Rs. 364 giving a total of Rs. 442 which is very near the rural average. The relatively higher average in respect of vehicles and tools in urban areas was due to the occurrence of larger number of motor transport operators compared with rural areas.

1.27. *Earnings :* Monthly gross earnings from transport enterprise per working household in rural areas was Rs. 44 and the corresponding figure for urban areas was Rs. 89. These estimates include only receipts in cash or kind for transport service rendered and exclude imputed value of own use which may be quite high.

1.28. Monthly net earnings (gross of depreciations of fixed assets) per working household in rural areas was Rs. 20 which is only 45 per cent of the gross earnings while in urban areas the figure was 62 per cent of gross earnings at Rs. 55.

CHAPTER TWO

EMPLOYMENT

2.1. The total number of persons working in household transport enterprise during October 1953 to March 1954 is estimated at 1.11 million for the country as a whole. According to the 1951 Census, the total working population was 142.2 million. Compared with this total in all occupations the proportion in household transport enterprise is very small viz., 0.78 per cent. When the increase in total working population from 1951 to 1953 is considered, the proprotion will, perhaps, be still lower. Figures separately for rural and urban areas are given in Table (2.1).

TABLE (2.1): ESTIMATED NUMBER AND PERCENTAGE OF PERSONS ENGAGED IN HOUSEHOLD TRANSPORT ENTERPRISE IN RURAL AND URBAN AREAS : OCTOBER 1953–MARCH 1954

area	estimated number working in household transport enterprise (000)	total working population from 1951 Census (000)	percentage working in household transport enterprise
(1)	(2)	(3)	(4)
1. rural	670	120,800	0.55
2. urban	444	21,400	2.07
3. total	1114	142,200	0.78

2.2. In rural areas, number of persons working in household transport enterprise was 670,000 which is only 0.55 per cent of total rural working population of 120.8 million while in urban areas the figure was 444,000 which is 2.07 per cent of total urban working population of 21.4 million.

2.3. The total number of households engaged in transport enterprise as the principal or subsidiary means of livelihood is 1.46 million as estimated from the seventh round of the survey. However, there were seasonal fluctuations and all these 1.46 million households were not actively engaged in the enterprise in the period of the survey. The number of households actually operating in the period is estimated at 0.91 million which is only 62.60 per cent of the total 1.46 million. This percentage was widely different for rural and urban areas as can be seen from Table (2.2).

2.4. Since 546,000 households were not actively engaged in rendering transport services during the reference period, the total number of persons working in household transport enterprise were in the remaining 914,000 households which were working during the reference period. Thus, for the country as a whole the average number of persons working in household transport enterprise per household (working in the reference period) during October 1953–March 1954 was 1.22. Table (2.3) shows that there is very little difference between rural and urban areas in regard to this average.

TABLE (2.2): ESTIMATED NUMBER OF HOUSEHOLDS ENGAGED IN HOUSEHOLD TRANSPORT ENTERPRISE IN RURAL AND URBAN AREAS: OCTOBER 1953–MARCH 1954

area	number of sample households	estimated number of households in transport enterprise (000) in the reference period			col. (3) as percentage of col. (5)
		working	not working	total	
(1)	(2)	(3)	(4)	(5)	(6)
1. rural	2742	546	515	1061	51.46
2. urban	1652	368	31	399	92.23
3. total	4394	914	546	1460	62.60

TABLE (2.3): ESTIMATED NUMBER OF PERSONS ENGAGED IN HOUSEHOLD TRANSPORT ENTERPRISE PER WORKING HOUSEHOLD: OCTOBER 1953–MARCH 1954.

area	number of sample households	estimated number of		
		working households (000)	persons in transport enterprise	
			total	per working household
(1)	(2)	(3)	(4)	(5)
1. rural	2742	546	670	1.23
2. urban	1652	368	444	1.21
3. total	4394	914	1114	1.22

2.5. The number of working persons included both household and hired labour. The data relating to household and hired labour were collected in terms of man-days per working day in the month preceding the date of interview. These estimates of household and hired labour in terms of man-days per working day for rural and urban areas are given in Table (2.4).

TABLE (2.4): HOUSEHOLD AND HIRED LABOUR IN TERMS OF MAN-DAYS PER WORKING DAY IN RURAL AND URBAN AREAS: OCTOBER 1953-MARCH 1954

area	number of sample households	man-days (000) per working day			hired man-days as per cent of total man-days
		household labour	hired labour	total	
(1)	(2)	(3)	(4)	(5)	(6)
1. rural	2742	564	72	636	11.32
2. urban	1652	377	51	428	11.92
3. total	4394	941	123	1064	11.56

2.6. Figures of man-days per working day show how many persons were working on an average on a working day. A small proportion of persons was not working full time in this line. This is indicated by the fact that while the average number of working persons for the country was 1,114,000 persons, the average man-days per working day were 1,064,000. Of the total man-days of 1,064,000 hired labour accounted for only 123,000, i.e., the percentage of hired man-days was only 11.56. This shows that small-scale transport services operated mostly with household labour. These percentages for the rural and urban areas were similar.

2.7. The monthly amounts disbursed as charges to hired labour are shown in Table (2.5).

TABLE (2.5): MONTHLY CHARGES FOR HIRED LABOUR IN RURAL AND URBAN AREAS: OCTOBER 1953-MARCH 1954

area	number of sample households	monthly charges for hired labour Rs. (000)	total man-days of hired labour per working day (000)	monthly charges for a hired labourer (Rs.)
(1)	(2)	(3)	(4)	(5)
1. rural	2742	2336	72	32.44
2. urban	1652	2444	51	47.92
3. total	4394	4780	123	38.86

2.8. The monthly charge for a hired labourer was Rs. 32.44 in rural areas as compared with Rs. 47.92 in urban areas. The average for the country as a whole was Rs. 38.86 per person.

2.9. The relative importance of the different types of transport in the rural areas from employment point of view is shown in Table (2.6).

2.10. Out of the total of 670,000 persons, 425,000 or 63.43 per cent were offering transport services with bullock or buffalo-drawn carts, 96,000 persons or 14.33 per cent with carts drawn by other animals or with pack animals. Boats and other water transports accounted for 81,000 persons or 12.09 per cent of total. In regard to the other two types given in Table (2.6) the estimates are 42 and 26 thousand which together contributed only 10.15 per cent of the total employment. The category 'others' includes man-driven transports like rickshaws, palanquins and dulies and power-driven transports like motor cars.

2.11. The position in respect of employment in the urban areas was different as can be seen from Table (2.7).

2.12. Out of the total of 444,000 persons, 63.52 per cent or 282,000 were engaged as rickshaw pullers, carters, porters etc. Employment with animal-drawn carts and vehicles accounted for 22.97 per cent and with pack animals 3.15 per cent.

Motor cars and vehicles accounted for 7.66 per cent. Employment in all other types of transports taken together came to 12,000 persons or 2.70 per cent of the total. The category 'others' includes water-borne transports like boats.

TABLE (2.6): NUMBER OF WORKING PERSONS BY TYPES OF TRANSPORT IN RURAL AREAS: OCTOBER 1953–MARCH 1954

type of transport	estimated number of working persons (000)	percentage of total
(1)	(2)	(3)
1. bullock and buffalo carts	425	63.43
2. pack animals and vehicles drawn by other animals	96	14.33
3. boats and other water transports	81	12.09
4. porters and coolies	42	6.27
5. others	26	3.88
6. total	670	100.00
7. number of sample households	2742	
8. number of sample villages	959	

TABLE (2.7): NUMBER OF WORKING PERSONS BY TYPES OF TRANSPORT IN URBAN AREAS: OCTOBER 1953–MARCH 1954

type of transport	estimated number of working persons (000)	percentage of total
(1)	(2)	(3)
1. man-drawn transports like push carts, cycle-rickshaws etc.	154	34.69
2. porters and coolies	128	28.83
3. animal-drawn carts and vehicles	102	22.97
4. motor cars and vehicles	34	7.66
5. pack animals	14	3.15
6. others	12	2.70
7. total	444	100.00
8. number of sample households	1652	
9. number of sample blocks	443	

COMPARISON WITH CENSUS 1951

2.13. The number of persons engaged in transport activities by various classifications have been shown in the occupational tables of the 1951 Population Census. These figures refer to all kinds of transports, such as by road, water, air and railways and to persons of all occupational status in both large- and small-scale enterprises. Besides, in the Census, the number of persons comprised all those who normally derived some income from transport enterprise while in this report, the number of working persons is based only on those households which were operating transport enterprise during the reference period of the survey. Furthermore, the Census figures refer to 1951 whereas the figures of the present survey refer to the period October 1953 to March 1954. The Census and the survey figures are thus not strictly comparable. Nevertheless, if the number of persons engaged in small-scale transport by road and water are separated from the total Census figures, a comparison of the figures thus arrived at with those of the present survey will be of interest.

2.14. The number of persons engaged in transport as principal means of livelihood are given in Table BIII in 1951 Census Report, Vol. I, Part IIC. A summary of these figures is shown in Table (2.8) arranged by rural and urban areas, by types of transport and by occupational status of the persons.

TABLE (2.8): NUMBER OF PERSONS ENGAGED IN TRANSPORT ENTERPRISE AS PRINCIPAL MEANS OF LIVELIHOOD : CENSUS 1951

	transport by	employers	employees	independent workers	total
	(1)	(2)	(3)	(4)	(5)
			rural		
1.	road	6,750	95,021	149,891	251,662
2.	water	2,557	38,348	29,975	70,880
3.	air	49	3,408	270	3,725
4.	railways	208	188,736	8,749	197,693
			urban		
1.	road	23,265	293,782	324,750	641,797
2.	water	3,321	120,558	22,009	145,888
3.	air	86	8,498	1,185	9,769
4.	railways	99	360,568	14,359	375,026

2.15. The number of persons who had transport activities as their secondary means of livelihood are given in Table BII in 1951 Census Report, Vol. I, Part IIB. Breakdowns by rural and urban areas, by transport type or by occupational status were not, however, given. A summary of these figures is given in Table (2.9).

TABLE (2.9): NUMBER OF PERSONS ENGAGED IN TRANSPORT ENTERPRISE AS SECONDARY MEANS OF LIVELIHOOD : CENSUS 1951

items	number of persons (earners and earning dependants)
(1)	(2)
1. persons with agriculture as principal means	406,218
2. persons with occupations other than agriculture as principal means	161,928
3. total (rural and urban)	568,146

2.16. The breakdown for rural and urban areas not being given, a division can be made on some reasonable assumptions. It will not be wrong to assume that all the persons who had agriculture as their principal means of livelihood were living in villages and that half of the persons who had non-agricultural occupations as their principal means of livelihood were also living in villages. The allocation of half to the villages in the latter case has been based on the fact that about half of the total persons who had non-agricultural occupations, with or without secondary occupations, were living in rural areas viz., 16.5 out of a total 32.4 million. Thus, persons who had transport operations as their secondary means of livelihood may be divided between rural and urban areas as follows : rural areas, 487,000 persons; and urban areas, 81,000 persons. These figures include employees too. Subtracting the number of employees calculated on the assumption that the proportion of employees in the total number of persons is the same in both the cases where transport activities were the principal or secondary means of livelihood, the number of employers and independent workers having transport services as a secondary means of livelihood came to 286,000 and 38,000 for the rural and urban areas respectively.

2.17. We may now take note of persons who may be considered as small-scale transport operators. In keeping with the definition of transport operators used in the sample survey under discussion, the figures of independent workers and employers in relation to transports by road and water in Table (2.8) will roughly show the number of persons engaged in small-scale operations as principal means of livelihood. In the present context the figures relating to air transport and railways are excluded. When the relevant figures in Table (2.8) are added, the number of persons for rural areas comes to 189,173 and that for urban areas 373,345. Taking principal and secondary means of livelihood together, the number of persons engaged in household transport operations (excluding employees) come to 475,000 and 411,000 for rural and urban areas respectively giving a total of 886,000.

2.18. These figures do not include employees in household transport enterprise and as such cannot be compared with the total estimates of number of persons given in Table (2.1). Even though separate estimate of persons other than employees is not directly available from the survey, a good approximation to this can be had by taking the household part of the estimated total number of persons based on proportion of household man-days to the total man-days (cf. Table 2.4). These adjusted figures for persons other than employees from the survey come to 594,000 and 391,000 for rural and urban areas respectively.

2.19. Now, a comparison between the two sets of figures of number of persons other than employees engaged in household transport enterprise may be made.

TABLE (2.10): A COMPARISON OF NUMBER OF PERSONS OTHER THAN EMPLOYEES ENGAGED IN HOUSEHOLD TRANSPORT ENTERPRISE AS OBTAINED FROM 1951 CENSUS AND NSS SEVENTH ROUND

area	number of persons other than employees in household transport enterprise (000)	
	Census 1951	NSS : Oct. 1953–March 1954
(1)	(2)	(3)
1. rural	475	594
2. urban	411	391
3. total	886	985

2.20. As mentioned in para 2.13 the figures are only roughly comparable since the gap between the two periods is nearly three years. The assumptions made in arriving at the figures are also to be noted. However, the table shows that about a million persons in the country were self-employed in transport service.

Chapter Three

FIXED CAPITAL

3.1. Data regarding fixed capital could be collected from all sample households engaged in transport activities irrespective of whether they were operating in the period of the survey or not. The value of fixed capital assets does not depend on the operation or non-operation of the household concerned. The per household and other figures in this chapter are therefore based on all households engaged in transport services. It may also be mentioned that the data collected in respect of fixed capital pertained to cost of acquisition of non-inherited assets only.

3.2. It may be of interest to know the extent to which inherited assets have affected the estimates. Fixed assets comprise land, buildings, draught and pack animals, and transport vehicles, tools etc. Of these, animals, vehicles, tools etc. do not present any difficulty in regard to the question of inheritance because the relative numbers were insignificant. In the case of land and buildings, however, the proportions were not insignificant and Table (3.1) gives the percentages of sample households which inherited land and buildings used in transport enterprise.

TABLE (3.1): PERCENTAGE OF SAMPLE HOUSEHOLDS WHICH INHERITED LAND AND BUILDINGS USED IN TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

area	number of sample households	percentage of households inheriting	
		land	building
(1)	(2)	(3)	(4)
1. rural	2742	21.52	6.60
2. urban	1652	3.69	2.48
3. total	4394	12.18	5.05

3.3. The proportions of households which inherited land were 21.52 and 3.69 per cent in the rural and urban areas respectively. The corresponding percentages for buildings are 6.60 and 2.48. The proportions were higher in rural areas with respect to both land and buildings. In both the areas the percentages were higher for land than for buildings.

3.4. A point in regard to estimates of fixed capital per household is that the figures may be vitiated to some extent by the inherited assets. In estimating totals of households, the inheriting households were also included even though no entries could be made regarding cost of acquisition of the fixed assets in most of these cases. In view of this, the average value of fixed capital per household is likely to be underestimated.

3.5. The values of land and buildings separately for rural and urban are as are given in Table (3.2) arranged by types of transport.

TABLE (3.2) : VALUES OF LAND AND BUILDINGS PER HOUSEHOLD USED IN TRANSPORT ENTERPRISE IN RURAL AREAS : OCTOBER 1953–MARCH 1954

type of transport	number of households engaged (000)	value per household in Rs.	
		land	buildings
(1)	(2)	(3)	(4)
1. bullock and buffalo carts	783	4.66	43.90
2. pack animals and vehicles drawn by other animals	131	3.59	50.24
3. boats and other water transports	42	0.10	1.71
4. porters and coolies	61	0.00	0.00
5. others	44	3.04	15.18
6. total	1061	4.01	39.30
7. number of sample households	2742		
8. number of sample villages	959		

3.6. It will be seen that the average value of land per household used only for this enterprise was very low. The average values of buildings were higher, e.g., Rs.43.90 in the case of bullock or buffalo carts and Rs.50.24 in the case of carts drawn by other animals or pack animals. Actually in a large proportion of cases no values were noted for land and buildings because these were not specifically assigned to carts and animals. It may be noted here that the field staff were instructed to allocate the value of land or buildings in proportion to the gross earnings wherever a household used them for more than one enterprise.

3.7. The figures for the urban areas are shown in Table (3.3).

TABLE (3.3) : VALUES OF LAND AND BUILDINGS PER HOUSEHOLD USED IN TRANSPORT ENTERPRISE IN URBAN AREAS : OCTOBER 1953–MARCH 1954

type of transport	estimated number of households engaged (000)	value per household in Rs.	
		land	buildings
(1)	(2)	(3)	(4)
1. motor cars and vehicles	12	96.00	308.75
2. man-drawn transport like push carts, cycle-rickshaws etc.	144	0.13	1.24
3. porters and coolies	120	0.00	0.00
4. animal-drawn carts and vehicles	107	14.74	47.78
5. pack animals	10	0.60	53.50
6. others	6	0.00	4.00
7. total	399	6.90	23.95
8. number of sample households	1652		
9. number of sample blocks	443		

3.8. In the urban areas, the average values of land used per household were much higher in comparison with those for rural areas, e.g., the figure for motor cars and vehicles was Rs. 96.00 and for animal-drawn carts Rs.14.74. For other types of transports the values of land used were insignificant. The value of buildings per household was as high as Rs. 308.75 for motor cars and vehicles. But the value for carts and animals was around Rs. 50 and was in level with similar values in rural areas. The values of buildings for other types of transport were insignificant.

3.9. The values of (a) draught or pack animals and (b) vehicles and tools, used for transport enterprise in the rural areas are given in Table (3.4) by types of transport.

TABLE (3.4): VALUES OF ANIMALS, VEHICLES AND TOOLS PER HOUSEHOLD USED IN TRANSPORT ENTERPRISE IN RURAL AREAS: OCTOBER 1953–MARCH 1954

type of transport	estimated number of households engaged (000)	value per household in Rs.	
		animals	vehicles and tools
(1)	(2)	(3)	(4)
1. bullock and buffalo carts	783	331.71	134.01
2. pack animals and vehicles drawn by other animals	131	246.53	42.75
3. boats and other water transports	42	—	815.57
4. porters and coolies	61	—	4.74
5. others	44	—	589.64
6. total	1061	275.23	161.19
7. number of sample households	2742		
8. number of sample villages	959		

3.10. The average value for animals per household was Rs. 331.71 in the case of bullock and buffalo carts, and Rs. 246.53 for other draught animals and pack animals. The question of animals does not arise in other cases. In regard to vehicles and tools, the per household value for boats and other water transports was the highest, namely, Rs. 815.57. 'Others' included motor vehicles also and hence the per household average came to a figure as high as Rs. 589.64.

3.11. Similar figures for the urban areas are given in Table (3.5).

3.12. The average value for animals per household was Rs. 249.76 for bullocks, buffaloes and other draught animals. The per household value was higher for the pack animals, namely, Rs.431.30. The figures for animals in the rural and urban tables are not entirely comparable because of their grouping in different ways. In regard to vehicles and tools, the per household value for motor cars and vehicles was Rs. 9242.75, for animal-drawn carts and vehicles Rs. 173.25, and for cycle-rickshaws and rickshaws Rs. 73.00. The accessories in relation to pack animals came to Rs. 19.20 per household, 'others' included boats and other water transports.

TABLE (3.5) : VALUES OF ANIMALS, VEHICLES AND TOOLS PER HOUSEHOLD USED IN TRANSPORT ENTERPRISE IN URBAN AREAS : OCTOBER 1953–MARCH 1954

type of transport	estimated number of households engaged (000)	value per household in Rs.	
		animals	vehicles and tools
(1)	(2)	(3)	(4)
1. motor cars and vehicles	12	—	9242.75
2. man-drawn transports like push carts, cyclo-rickshaws etc.	144	—	73.00
3. porters and coolies	120	—	0.20
4. animal-drawn carts and vehicles	107	249.76	173.25
5. pack animals	10	431.30	19.20
6. others	6	—	816.17
7. total	399	77.79	363.60
8. number of sample households	1652		
9. number of sample blocks	443		

3.13. The figures of the main items of the fixed capital are now brought together for convenience in one table, namely, Table (3.6).

TABLE (3.6) : BREAKDOWNS OF FIXED CAPITAL IN VALUES AND PERCENTAGES FOR DIFFERENT TYPES OF TRANSPORT IN RURAL AREAS : OCTOBER 1953– MARCH 1954

type of transport	land	buildings	animals	vehicles and tools	total
(1)	(2)	(3)	(4)	(5)	(6)
(i) value per household in Rs.					
1. bullock and buffalo carts	4.66	43.90	331.71	134.01	514.28
2. pack animals and vehicles drawn by other animals	3.59	50.24	246.53	42.75	343.11
3. boats and other water transports	0.10	1.71	—	815.57	817.38
4. all types together	4.01	39.30	275.23	161.19	479.64
(ii) percentage distribution					
5. bullock and buffalo carts	0.91	8.54	64.50	26.05	100.00
6. pack animals and vehicles drawn by other animals	1.05	14.64	71.85	12.46	100.00
7. boats and other water transports	0.01	0.21	—	99.78	100.00
8. all types together	0.84	8.17	57.38	33.61	100.00
9. number of sample households	2742				
10. number of sample villages	959				

3.14. The breakdowns for bullock and buffalo carts show that roughly two-thirds of the fixed capital went for the draught animals, about a fourth for vehicles and tools, and about a tenth for land and buildings. For the next group, since the pack animals have been combined with vehicles drawn by animals other than bullocks and buffaloes, the proportion of value for animals came higher to nearly three-fourths and that for vehicles and tools to only about a tenth of the total value of the fixed capital. In the case of boats and other water transports, almost the whole value was accounted for by the craft.

3.15. Similar figures for the urban areas are given in Table (3.7).

TABLE (3.7): BREAKDOWNS OF FIXED CAPITAL IN VALUES AND PERCENTAGES FOR DIFFERENT TYPES OF TRANSPORT IN URBAN AREAS: OCTOBER 1953–MARCH 1954

(i) value per household in Rs.

type of transport	land	buildings	animals	vehicles and tools	total
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	96.00	308.75	—	9242.75	9647.50
2. animal-drawn carts and vehicles	14.74	47.78	249.76	173.25	485.53
3. man-drawn transports like push carts, cycle-rickshaws etc.	0.13	1.24	—	73.00	74.37
4. all types together	6.90	23.95	77.79	363.60	472.24
(ii) percentage distribution					
5. motor cars and vehicles	1.00	3.20	—	95.80	100.00
6. animal-drawn carts and vehicles	3.04	9.84	51.44	35.68	100.00
7. man-drawn transports like push carts, cycle-rickshaws etc.	0.17	1.67	—	98.16	100.00
8. all types together	1.46	5.07	16.47	77.00	100.00
9. number of sample households	1652				
10. number of sample blocks	443				

3.16. It will be seen that although the per household values of land and buildings used in the case of motor cars and vehicles were much higher than those for other types of transport, their proportion to the total value of fixed capital was only 4.20 per cent. Vehicles and tools accounted for as much as 95.80 per cent. The percentage breakdowns in the case of animal-drawn carts and vehicles were 35.68 per cent for vehicles and tools, 51.44 per cent for draught animals and 12.88 per cent for land and buildings used. As for the type of transport pulled by man, nearly the whole value went for vehicles and tools.

CHAPTER FOUR

COSTS AND EARNINGS

RURAL AREAS

4.1. The particulars on costs and earnings included in this chapter refer to operating households only. The households which were not operating at the time of the survey had very little to show in regard to running costs and earnings and in most cases the entries in the schedules were just 'nil'. Figures under these items were meaningful only for the operating households and only a monthly picture of costs and earnings during the period, October 1953 to March 1954, will be obtained from the data. Throughout the discussion in this chapter, by costs the input is meant.

4.2. The different costs arranged by types of transport are given in Table (4.1).

TABLE (4.1): BREAKDOWNS OF MONTHLY COSTS PER OPERATING HOUSEHOLD IN TRANSPORT ENTERPRISE IN RURAL AREAS: OCTOBER 1953–MARCH 1954

type of transport	estimated number of households (000)	costs per household (Rs.)				
		fuel and lubricants	animal feed	repair and maintenance	other expenses	total monthly costs
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. bullocks and buffalo carts	375	0.54	23.22	11.09	0.11	34.96
2. carts and vehicles drawn by other animals and pack animals	89	0.32	29.46	13.30	0.23	43.31
3. boats and other water transports	31	12.39	—	13.77	0.19	26.35
4. porters and coolies	35	—	—	0.03	0.03	0.06
5. others	16	74.62	—	68.06	2.56	145.24
6. all transports	546	3.32	20.75	12.56	0.20	36.83
7. number of sample households	2742					
8. number of sample villages	959					

4.3. The total recurring costs per month per household were Rs.34.96 for bullock and buffalo carts, Rs. 43.31 for carts and vehicles drawn by other animals and pack animals, Rs.26.35 for boats and other water transports and Re. 0.06 for porters and coolies. The figure for 'others', which included motor cars and vehicles, was Rs. 145.24.

4.4. As regards breakdowns of costs, it may be seen that the per household costs for fuel, lubricants etc., were the highest for 'others' which included motor cars and vehicles namely, Rs.74.62. The cost was Rs.12.39 for boats and other water transports. For all types of animal-drawn carts and vehicles these costs were negligible. Costs on cattle feed were relevant for two types of transport only, namely, bullock and buffalo carts, and carts and vehicles drawn by other animals and pack animals. The repair and maintenance costs for bullock and buffalo carts, other animal transports and for boats and other water transports were of the same order varying between Rs.11.09 and Rs.13.77. There was very little by way of repair and maintenance charges for porters and coolies, but for 'others' which included motor vehicles these costs were high, namely, Rs.68.06. Other expenses, which included charges for stores and services, were of low order in all cases.

4.5. The costs for animal feed raise two special problems. Firstly, in addition to costs of fodder which was purchased the amount also included imputed value of supplies obtained from one's own farm or otherwise. The question is, to what extent the inclusion of such imputed value will give a realistic picture of costs. Secondly, most of these animals, particularly bullocks and buffaloes, were not used exclusively for transport operations but for other work as well including cultivation. What proportion of costs of animal feed should we allocate to transport operations ?

4.6. Let us first consider the question of home supply. Of the amount of Rs.23.22 per household for feed for bullocks and buffaloes, Rs.11.49 or about 50 per cent represented value imputed to fodder supply obtained in other ways than by purchase. In the case of other animals, the proportion of home supply was much lower namely, Rs.6.34 out of Rs.29.46 or just above a fifth. A part of the fodder came from the produce of a farm which might have alternative use but a part also came through straying.

4.7. As for the question of allocation, since the animals were used for more than one purpose, the costs of feed should be allocated over different activities proportionately. However, as the particulars on which allocations would be based were not collected in the schedule, an idea of the proportionate use of bullocks and buffaloes for transport work was obtained from another schedule used in the same survey. The latter schedule was canvassed to collect data on agriculture and animal husbandry wherein information on different uses of farm animals over a whole year was recorded. The particulars of those sample households who were also engaged in transport services showed that 17.3 per cent of their cattle time was used in transport work. This figure, however, referred to a whole year and not just to the period, October 1953 to March 1954, concerned in the present survey. But since we have no data on allocation other than this, it is being used here to get the suitable cost of feed for bullocks and buffaloes. The figure should be considered as approximate only and when further information on this point is collected, necessary adjustments will be made. The allocated cost of cattle feed thus comes to Rs.4.02 per household for transport operations (17.3 per cent of Rs.23.22). Similar calculations for feed of other animals could not be made due to want of suitable data.

4.8. The monthly earnings of the households by different types of transport are shown in Table (4.2).[1]

TABLE (4.2) : GROSS EARNINGS PER MONTH PER HOUSEHOLD BY GOODS AND PASSENGER TRAFFIC IN RURAL AREAS : OCTOBER 1953–MARCH 1954

type of transport	estimated number of households (000)	gross earnings per month per household				
		amount (Rs.)			percentage breakdown	
		goods traffic	passenger traffic	total	goods traffic	passenger traffic
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. bullock and buffalo carts	375	22.98	2.88	25.86	88.86	11.14
2. pack animals and carts and vehicles drawn by other animals	89	19.01	28.07	47.08	40.38	59.62
3. boats and other water transports	31	150.39	7.00	157.39	95.55	4.45
4. porters and coolies	35	22.37	—	22.37	100.00	—
5. others	16	160.75	122.44	283.19	56.76	43.24
6. all transports	546	33.56	10.54	44.10	76.10	23.90
7. number of sample households	2742					
8. number of sample villages	959					

4.9. On the whole, it will be seen, the major part of the earnings came from goods traffic. As compared with the total from both the sources, the proportions of earnings from goods traffic varied between 89 to 100 per cent for bullock and buffalo carts, water transport, and porters and coolies. In the case of carts and vehicles drawn by other animals and pack animals, the proportion of earnings from goods traffic was 40.38 per cent and that from passenger traffic 59.62 per cent. The proportion of earnings from goods traffic was 56.76 per cent in the case of 'others' which included motor vehicles, and that from passenger traffic was 43.24 per cent.

4.10. A comparison of the earnings and costs by different types of transport will be of interest and is given in Table (4.3).

4.11. The differences between the gross earnings and total costs are not exactly net incomes because these are gross of depreciation of fixed assets. The approximate nature of costs for bullock and buffalo carts, and for transports by other animals, due to difficulties of allocation should also be noted. However, the figures will show the order of net incomes. It will be seen that the per household incomes for the first two types of transports namely, transport by all kinds of animals, were low as compared with the last three types of transport. The reason for this is that most of the households offering the first two types of service worked only partially for transport operations.

---

[1] In the National Sample Survey, the general approach to all enterprise activities is to get the values of production both sold out and consumed by the households. But this may not be rigidly stuck to in the service type of enterprises because of the very nature of the enterprise.

TABLE (4.3) : MONTHLY EARNINGS AND COSTS PER HOUSEHOLD IN TRANSPORT ENTERPRISE IN RURAL AREAS : OCTOBER 1953–MARCH 1954

type of transport	estimated number of households (000)	per household per month (Rs.)		
		gross earnings	total costs*	difference
(1)	(2)	(3)	(4)	(5)
1. bullock and buffalo carts	375	25.86	15.76	10.10
2. pack animals and carts and vehicles drawn by other animals	89	47.08	43.31	3.77
3. boats and other water transports	31	157.39	26.35	131.04
4. porters and coolies	35	22.37	0.06	22.31
5. others	16	283.19	145.24	137.95
6. all transports	546	44.10	23.64	20.46
7. number of sample households	2742			
8. number of sample villages	959			

* cost of cattle feed allocated in proportion to use of the animals in transport operations

URBAN AREAS

4.12. As in the case of rural areas, the figures of costs and earnings in urban areas also relate to operating households only; and the figures give a monthly picture during the period October 1953 to March 1954. The breakdowns of the running costs are given in Table (4.4).

TABLE (4.4) : BREAKDOWNS OF MONTHLY COSTS PER OPERATING HOUSEHOLD IN TRANSPORT ENTERPRISE IN URBAN AREAS : OCTOBER 1953–MARCH 1954

type of transport	estimated number of households (000)	cost per household in Rs.				
		fuel, lubricants	cattle feed	repair and maintenance	other expenses	total monthly costs
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. motor cars and vehicles	10	422.70	—	198.60	29.60	650.90
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	0.38	—	2.60	0.07	3.05
3. porters and coolies	117	—	—	—	0.01	0.01
4. animal-drawn carts and vehicles	88	0.87	45.44	11.25	0.32	57.88
5. pack animals	8	—	38.75	15.38	1.25	55.38
6. others	5	0.80	—	8.20	0.60	9.60
7. all transports	368	11.85	11.71	9.52	0.95	34.03
8. number of sample households	1652					
9. number of sample blocks	443					

4.13. The per household total of monthly costs was the highest for motor cars and vehicles, namely, Rs.650.90. The expenses for all the other types of transport were much below this level. The costs of transports using animals were somewhat higher than the rest because of the expenses for maintaining the animals. For example, the total cost per household was Rs.57.88 per month for animal-drawn carts and vehicles and Rs.55.38 for pack animals as compared with Rs.3.05 for transport like push carts, cycle-rickshaws, rickshaws etc.

4.14. Table (4.5) shows the earnings from goods and passenger traffic for the different types of transport.

TABLE (4.5) : GROSS EARNINGS PER MONTH PER HOUSEHOLD BY GOODS AND PASSENGER TRAFFIC IN URBAN AREAS : OCTOBER 1953–MARCH 1954

type of transport	estimated number of households (000)	gross earnings per month per household				
		amount Rs.			percentage breakdown	
		goods traffic	passenger traffic	total	goods traffic	passenger traffic
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. motor cars and vehicles	10	578.60	620.10	1198.70	48.27	51.73
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	23.38	35.69	59.07	39.58	60.42
3. porters and coolies	117	34.53	—	34.53	100.00	—
4. animal-drawn carts and vehicles	88	50.98	34.61	85.59	59.56	40.44
5. pack animals	8	82.38	0.50	82.88	99.40	0.60
6. others	5	60.40	5.80	66.20	91.24	8.76
7. all transports	368	50.39	38.80	89.19	56.50	43.50
8. number of sample households	1652					
9. number of sample blocks	443					

4.15. The income from passenger traffic was 43.50 per cent of the total and this was much higher than the proportion for the rural areas which was 23.90 per cent. The earnings were roughly evenly divided between the two sources, goods and passenger traffic, in the case of motor cars and vehicles, man-drawn transports, and animal-drawn transports, varying within 40 to 60 per cent. The other means of transport namely, porter and coolies, pack animals and others got their earnings almost wholly from goods traffic.

4.16. The earnings and costs of the household transport operations in urban areas are shown in Table (4.6).

4.17. As mentioned earlier in relation to the figures of the rural areas, the differences between the earnings and costs are gross of depreciation of fixed assets and to that extent the figures are not net incomes. As we would expect, the per household income was the highest for motor cars and vehicles namely, Rs.547.80. The incomes of households using other types of transport were very much lower and this was particularly so for animal-drawn carts and vehicles. For the same type of transport, the level of income was, however, higher in urban areas than in rural areas.

TABLE (4.6): MONTHLY EARNINGS AND COSTS PER HOUSEHOLD IN TRANSPORT ENTERPRISE IN URBAN AREAS: OCTOBER 1953–MARCH 1954

type of transport	estimated number of households (000)	per household per month (Rs.) gross earnings	total costs	difference
(1)	(2)	(3)	(4)	(5)
1. motor cars and vehicles	10	1198.70	650.90	547.80
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	59.07	3.05	56.02
3. porters and coolies	117	34.53	0.01	34.52
4. animal-drawn carts and vehicles	88	85.59	57.88	27.71
5. pack animals	8	82.88	55.38	27.50
6. others	5	66.20	9.60	56.60
7. all transports	368	89.19	34.03	55.16
8. number of sample households	1652			
9. number of sample blocks	443			

CHAPTER FIVE

DESCRIPTION OF TABLES

5.1. There are 12 tables in Appendix I presenting information on household transport enterprises in rural and urban areas of India for the period October 1953 to March 1954. The figures relate only to those households which were operating during the reference period of the survey. The figures for rural and urban areas have been shown separately. The classifications of transport types are different for the two areas because the predominating types were different.

5.2. The numbers of households given in column (2) of the tables are the estimated numbers of households engaged in respective types of transport enterprises and column (1) in the tables gives the descriptions of the different types of transport. The estimates of totals for different items are given in thousands in the tables and the per household estimates were arrived at by dividing the totals by the corresponding numbers of households. The number of sample villages or sample blocks, as the case may be, with the number of sample households are shown in the bottom-most lines of each table.

5.3. Table 1(1) gives the number of households operating the household transport enterprise in the survey period, and average number of persons engaged per working day. For the rural areas, the different types of transport have been classified as follows : (1) bullock and buffalo carts, (2) vehicles drawn by other animals and pack animals, (3) boats and other water transports, (4) porters and coolies and (5) other transports which included motor vehicles also. For the urban areas, the classifications are, (1) motor cars and vehicles, (2) man-drawn transports like push carts, cycle-rickshaws, rickshaws etc., (3) porters and coolies, (4) animal-drawn carts and vehicles, (5) pack animals and (6) other transports including water transports.

5.4. The total number of persons working in transport operations was 670,000 persons per working day in the rural areas and 444,000 in the urban areas. The breakdowns for the different types of transport are shown in column (3). The per household estimates are shown in column (4).

5.5. Table 1(2) shows monthly expenditures on fuel, lubricants and other similar materials as obtained by purchase and by home supply. In the rural areas, the overall per household expenditure on these materials as obtained by purchase and by home supply were Rs.3.05 and Re.0.26 respectively. Only the figures for boats and water transports and for 'others' which included motor vehicles are substantially high. In the urban areas, the overall per household expenditures were much higher because of greater use of motor vehicles.

5.6. Table 1(3) shows expenditures on cattle feed separately as obtained by purchase and home supply. The figures naturally relate to only those types of transport where animals were used.

5.7. Table 1(4) shows the percentage breakdowns of expenditures on fuel, lubricants etc., and cattle feed obtained by purchase and home supply. The percentages regarding fuel, lubricants etc., are given for purchase in column (2) and for home supply in column (3). 92.10 per cent of the total expenditure on fuel etc., in rural areas related to materials which were purchased and only 7.90 per cent to materials which were obtained from home supply. The breakdowns for cattle feed are of more interest and 56.99 per cent were obtained by purchase and 43.01 per cent by home supply. In the urban areas, the proportion obtained from home supply was very low, e.g., for fuel, lubricants etc., it was 0.05 per cent and for cattle feed it was 9.42 per cent.

5.8. Table 1(5) shows the totals of purchases and home supplies regarding expenditures on fuel, lubricants etc., and cattle feed. The overall per household monthly expenditure on fuel, lubricants, etc., was Rs.3.32 in rural areas and Rs.11.85 in urban areas. The household monthly expenses for cattle feed were Rs.20.75 in rural areas and Rs.11.71 in urban areas.

5.9. Table 1(6) relates to the monthly expenditures on repair and maintenance and rent in household transport enterprises. Rent included charges of hiring, if any, of vehicles and equipments. The per household monthly expenditures on repair and maintenance in rural and urban sectors were Rs.12.56 and Rs.9.52 respectively. The expenses on rent per household were Re.0.45 and Rs.7.64 in the rural and urban areas respectively.

5.10. Table 1(7) shows monthly expenditures on other stores and on services purchased. Table 1(8) shows monthly expenditure for licences, taxes and local rates; and interest on loan-capital in household transport enterprise.

5.11. Table 1(9) gives the household and hired man-days per working day in household transport enterprise in rural and urban areas, during the period October 1953 to March 1954. The per household figure for hands belonging to the household was 1.03 persons per working day in the rural areas and 1.02 persons in the urban areas. The number of hired man-days per working day in the rural areas was only 0.13 person per household and for urban areas 0.14 person.

5.12. Table 1(10) shows the monthly hired labour charges in household transport enterprise in rural and urban areas during the period October 1953 to March 1954. It will be seen that in rural areas, the hired labour charges were the highest for boat and other water transports, namely, Rs. 39.94 per household. In the urban areas, the highest charges were recorded for motor cars and vehicles, namely, Rs.174.40 per household.

5.13. Table 1(11) shows the monthly gross earnings from goods and passenger traffic in household transport enterprise in rural and urban areas for the period October 1953 to March 1954. In the rural areas, the per household earnings from goods traffic were Rs.33.56 and from passenger traffic Rs.10.54. In the urban areas, the average earnings from goods traffic were Rs. 50.39 per household and Rs.38.80. from passenger traffic.

5.14. Table 1(12) gives the monthly total of earnings from both the sources in household transport enterprise. Column (5) shows the monthly gross earning per household and column (6) the monthly gross earning per working person. In rural areas, monthly gross income per household was Rs.44.10 and the per working person figure was Rs.35.94. In urban areas, the corresponding figures were Rs.89.19 and Rs.73.93 respectively.

5.15. Appendix II, as stated in paras 1.7 and 1.8 of chapter one, consists of all tables giving the different estimates pertaining to all households offering transport services but not necessarily operating in the reference period of the survey. Figures of the fixed assets only have been discussed in the text of this report. The rest of the figures, especially those relating to other items, have not been discussed, because these figures with reference to operating households only have been meaningful in the context of this report. The figures relating to all households have, however, been given with the report for the reason that these will be useful for studying the seasonal variations when similar figures are obtained from subsequent surveys. The arrangement of columns in the tables is similar to those in Appendix I.

5.16. Table 2(1) shows in column (2) the number of households engaged in transport operations, at one time or another, in the course of a year preceding the date of survey. The households have been arranged by types of transport operated by them. Totals of persons and household sizes have been shown in columns (3) and (4). Column (5) gives the average number of persons per working day and column (6) the per household figures.

5.17. Table 2(2) relates to the values of land and buildings used in household transport enterprise in rural and urban areas. Column (3) shows the values of land in thousands of rupees, while column (4) shows the values of land per household. Similarly, columns (5) and (6) give respectively the total values of buildings used by households in transport enterprise and the values per household.

5.18. Table 2(3) shows the values of draught and pack animals and transport vehicles and tools used in household transport enterprise in rural and urban areas. The figures regarding animals are relevant to certain types of transport only.

5.19. Table 2(4) relates to monthly expenditure on fuel, lubricant and other materials as obtained by purchase and home supply in household transport enterprise in rural and urban areas. In the rural areas, the overall per household expenditures for supplies obtained by purchase and from home were Rs.1.63 and Re.0.15 respectively. In the urban areas, the corresponding values were Rs.11.36 and Re.0.01.

5.20. Table 2(5) shows monthly expenditures on cattle feed as obtained by purchase and home supply. Total expenditures on the animals have been shown and no allocations in proportion to the use of these animals in transport operations have been made here. The expenditures in this respect are naturally in regard to bullock- and buffalo-drawn carts, pack animals, and animal-driven transports only.

5.21. Table 2(6) shows the percentage breakdowns of expenses on fuel, lubricants etc., and on cattle feed arranged by purchase and home supply. The arrangement of the table is the same as Table 1(4) in Appendix I.

5.22. Table 2(7) shows the total values of purchase and home supply taken together in regard to fuel, lubricants etc., and cattle feed.

5.23. Table 2(8) shows the monthly expenses on repair and maintenance and on rent in the period of October 1953 to March 1954. The per household monthly expenditure on repair and maintenance in rural and urban areas were Rs.8.68 and Rs.9.28 respectively. The per household figures for rent were Re.0.28 for rural areas and Rs.7.21 for urban areas.

5.24. Table 2(9) gives the monthly expenditures on other stores and services and Table 2(10) the monthly expenses for licences, taxes and rates, and on interest on loan capital. The per household values were very low in both the areas.

5.25. Table 2(11) gives the household and hired man-days per working day in household transport enterprise in rural and urban areas. The totals of man-days worked in both the categories are the same as in Table 1(9) of Appendix I, but the per household figures here are different because the totals were divided by all households engaged in transport services.

5.26. Table 2(12) relates to the monthly hired labour charges in household transport enterprises. Here also the total figures are the same as in Table 1(10) in Appendix I, but the per household figures are different because of larger number of households.

5.27. Tables 2(13) and 2(14) show figures regarding monthly gross earnings from goods and passenger traffic in household transport enterprise and the per household and per capita values of these earnings. Since the earnings arose only in relation to the operating households, the values of total earnings were the same as those given in Tables 1(11) and 1(12) of Appendix I. But since these values were spread over all the households connected with transport activities and not confined only to households operating at that time, the per household and per capita values are smaller.

APPENDIX I

LIST OF TABLES

(OPERATING HOUSEHOLDS)

PAGE

TABLE 1(1) : Number of households operating in small-scale transport enterprise and average number of persons engaged per working day : October 1953–March 1954 55

TABLE 1(2) : Value of monthly consumption of fuel, lubricants etc., obtained by purchase and by home supply in household transport enterprise : October 1953–March 1954 56

TABLE 1(3) : Monthly value of cattle feed obtained by purchase and by home supply in household transport enterprise : October 1953–March 1954 57

TABLE 1(4) : Percentage breakdown of expenses on fuel, lubricants etc., and cattle feed obtained by purchase and by home supply in household transport enterprise : October 1953–March 1954 58

TABLE 1(5) : Total value of monthly consumption of fuel, lubricants etc., and cattle feed in household transport enterprise : October 1953–March 1954 59

TABLE 1(6) : Monthly expenditure on repair and maintenance and rent in household transport enterprise : October 1953–March 1954 60

TABLE 1(7) : Monthly expenditure on stores and services in household transport enterprise : October 1953–March 1954 61

TABLE 1(8) : Monthly expenditure on licences, taxes etc., and interest on loan capital in household transport enterprise : October 1953–March 1954 62

TABLE 1(9) : Household and hired man-days per working day in household transport enterprise : October 1953–March 1954 63

TABLE 1(10) : Monthly hired labour charges in household transport enterprise : October 1953–March 1954 64

TABLE 1(11) : Monthly gross earnings from goods and passenger traffic in household transport enterprise : October 1953–March 1954 65

TABLE 1(12) : Monthly gross earnings in household transport enterprise : October 1953–March 1954 66

APPENDIX II

(ALL HOUSEHOLDS IN TRANSPORT OPERATIONS)

TABLE 2(1) : Number of households engaged in small-scale transport enterprise, household size and average number of persons engaged per working day : October 1953–March 1954 67

PAGE

TABLE 2(2) : Value of land and buildings used in household transport enterprise : October 1953–March 1954 68

TABLE 2(3) : Value of draught and pack animals and vehicles and tools used in household transport enterprise : October 1953–March 1954 69

TABLE 2(4) : Value of monthly consumption of fuel, lubricants etc., obtained by purchase and by home supply in household transport enterprise : October 1953–March 1954 70

TABLE 2(5) : Monthly value of cattle feed obtained by purchase and by home supply in household transport enterprise : October 1953–March 1954 71

TABLE 2(6) : Percentage breakdown of expenses on fuel, lubricants etc., and cattle feed obtained by purchase and by home supply in household transport enterprise : October 1953–March 1954 72

TABLE 2(7) : Total value of monthly consumption of fuel, lubricants etc., and cattle feed in household transport enterprise : October 1953–March 1954 73

TABLE 2(8) : Monthly expenditure on repair and maintenance and rent in household transport enterprise : October 1953–March 1954 74

TABLE 2(9) : Monthly expenditure on stores and services in household transport enterprise : October 1953–March 1954 75

TABLE 2(10) : Monthly expenditure on licences, taxes etc., and interest on loan capital in household transport enterprise : October 1953–March 1954 76

TABLE 2(11) : Household and hired man-days per working day in household transport enterprise : October 1953–March 1954 77

TABLE 2(12) : Monthly hired labour charges in household transport enterprise : October 1953–March 1954 78

TABLE 2(13) : Monthly gross earnings from goods and passenger traffic in household transport enterprise : October 1953–March 1954 79

TABLE 2(14) : Monthly gross earnings in household transport enterprise : October 1953–March 1954 80

APPENDIX I

TABLE 1(1): NUMBER OF HOUSEHOLDS OPERATING IN SMALL-SCALE TRANSPORT ENTERPRISE AND AVERAGE NUMBER OF PERSONS ENGAGED PER WORKING DAY: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	number of persons per working day	
		total (000)	per household
(1)	(2)	(3)	(4)
1. bullock and buffalo carts	375	425	1.13
2. pack animals and vehicles drawn by other animals	89	96	1.08
3. boats and other water transports	31	81	2.61
4. porters and coolies	35	42	1.20
5. others	16	26	1.62
6. total	546	670	1.23
7. number of sample households	2742		
8. number of sample villages	959		

urban areas

type of transport	estimated number of households (000)	number of persons per working day	
		total (000)	per household
(1)	(2)	(3)	(4)
1. motor cars and vehicles	10	34	3.40
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	154	1.10
3. porters and coolies	117	128	1.09
4. animal-drawn carts and vehicles	88	102	1.16
5. pack animals	8	14	1.75
6. others	5	12	2.40
7. total	368	444	1.21
8. number of sample households	1652		
9. number of sample blocks	443		

TABLE 1(2): VALUE OF MONTHLY CONSUMPTION OF FUEL, LUBRICANTS ETC., OBTAINED BY PURCHASE AND BY HOME SUPPLY IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	by purchase		by home supply	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	199	0.53	4	0.01
2. pack animals and vehicles drawn by other animals	89	29	0.32	0	0.00
3. boats and other water transports	31	249	8.03	135	4.35
4. porters and coolies	35	0	0.00	0	0.00
5. others	16	1190	74.38	4	0.25
6. total	546	1667	3.05	143	0.26
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	by purchase		by home supply	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	4227	422.70	0	0.00
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	52	0.37	1	0.01
3. porters and coolies	117	0	0.00	0	0.00
4. animal-drawn carts and vehicles	88	76	0.86	1	0.01
5. pack animals	8	0	0.00	0	0.00
6. others	5	4	0.80	0	0.00
7. total	368	4359	11.85	2	0.01
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(3): MONTHLY VALUE OF CATTLE FEED OBTAINED BY PURCHASE AND BY HOME SUPPLY IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	cattle feed			
		by purchase		by home supply	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	4400	11.73	4310	11.49
2. pack animals and vehicles drawn by other animals	89	2058	23.12	564	6.34
3. boats and other water transports	31	0	0.00	0	0.00
4. porters and coolies	35	0	0.00	0	0.00
5. others	16	0	0.00	0	0.00
6. total	546	6458	11.83	4874	8.93
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	cattle feed			
		by purchase		by home supply	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	0	0.00	0	0.00
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	0	0.00	0	0.00
3. porters and coolies	117	0	0.00	0	0.00
4. animal-drawn carts and vehicles	88	3593	40.83	406	4.61
5. pack animals	8	310	38.75	0	0.00
6. others	5	0	0.00	0	0.00
7. total	368	3903	10.61	406	1.10
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(4): PERCENTAGE BREAKDOWN OF EXPENSES ON FUEL, LUBRICANTS ETC., AND CATTLE FEED OBTAINED BY PURCHASE AND BY HOME SUPPLY IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural area

type of transport	percentage breakdown					
	fuel, lubricants etc.			cattle feed		
	purchase	home supply	total	purchase	home supply	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. bullock and buffalo carts	98.03	1.97	100.00	50.52	49.48	100.00
2. pack animals and vehicles drawn by other animals	100.00	0.00	100.00	78.49	21.51	100.00
3. boats and other water transports	64.84	35.16	100.00	—	—	—
4. porters and coolies	—	—	—	—	—	—
5. others	99.66	0.34	100.00	—	—	—
6. total	92.10	7.90	100.00	56.99	43.01	100.00
7. number of sample households	2742					
8. number of sample villages	959					

urban areas

type of transport	percentage breakdown					
	fuel, lubricants etc.			cattle feed		
	purchase	home supply	total	purchase	home supply	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. motor cars and vehicles	100.00	0.00	100.00	—	—	—
2. man-drawn transports like push carts, cycle-rickshaws etc.	98.11	1.89	100.00	—	—	—
3. porters and coolies	—	—	—	—	—	—
4. animal-drawn carts and vehicles	98.70	1.30	100.00	89.85	10.15	100.00
5. pack animals	—	—	—	100.00	0.00	100.00
6. others	100.00	0.00	100.00	—	—	—
7. total	99.95	0.05	100.00	90.58	9.42	100.00
8. number of sample households	1652					
9. number of sample blocks	443					

TABLE 1(5): TOTAL VALUE OF MONTHLY CONSUMPTION OF FUEL, LUBRICANTS ETC., AND CATTLE FEED IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	fuel, lubricants etc.		cattle feed	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	203	0.54	8710	23.22
2. pack animals and vehicles drawn by other animals	89	29	0.32	2622	29.46
3. boats and other water transports	31	384	12.39	0	0.00
4. porters and coolies	35	0	0.00	0	0.00
5. others	16	1194	74.62	0	0.00
6. total	546	1810	3.32	11332	20.75
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	fuel, lubricants etc.		cattle feed	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	4227	422.70	0	0.00
2. man-drawn transports like push carts, cycle-richshaws etc.	140	53	0.38	0	0.00
3. porters and coolies	117	0	0.00	0	0.00
4. animal-drawn carts and vehicles	88	77	0.87	3999	45.44
5. pack animals	8	0	0.00	310	38.75
6. others	5	4	0.80	0	0.00
7. total	368	4361	11.85	4309	11.71
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(6): MONTHLY EXPENDITURE ON REPAIR AND MAINTENANCE AND RENT IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953-MARCH 1954

rural areas

type of transport	estimated number of households (000)	repair and maintenance		rent	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	4157	11.09	89	0.24
2. pack animals and vehicles drawn by other animals	89	1184	13.30	6	0.07
3. boats and other water transports	31	427	13.77	80	2.58
4. porters and coolies	35	1	0.03	0	0.00
5. others	16	1089	68.06	71	4.44
6. total	546	6858	12.56	246	0.45
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	repair and maintenance		rent	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	1986	198.60	292	29.20
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	364	2.60	2175	15.54
3. porters and coolies	117	0	0.00	8	0.07
4. animal-drawn carts and vehicles	88	990	11.25	334	3.80
5. pack animals	8	123	15.38	2	0.25
6. others	5	41	8.20	1	0.20
7. total	368	3504	9.52	2812	7.64
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(7): MONTHLY EXPENDITURE ON STORES AND SERVICES IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	other stores		services	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	24	0.06	18	0.05
2. pack animals and vehicles drawn by other animals	89	2	0.02	19	0.21
3. boats and other water transports	31	1	0.03	5	0.16
4. porters and coolies	35	0	0.00	1	0.03
5. others	16	8	0.50	33	2.06
6. total	546	35	0.06	76	0.14
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	other stores		services	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	156	15.60	140	14.00
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	1	0.01	9	0.06
3. porters and coolies	117	0	0.00	1	0.01
4. animal-drawn carts and vehicles	88	5	0.06	23	0.26
5. pack animals	8	10	1.25	0	0.00
6. others	5	0	0.00	3	0.60
7. total	368	172	0.47	176	0.48
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(8) : MONTHLY EXPENDITURE ON LICENCES, TAXES ETC., AND INTEREST ON LOAN CAPITAL IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953-MARCH 1954

rural areas

type of transport	estimated number of households (000)	licences, taxes and rates		interest on loan capital	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	51	0.14	100	0.27
2. pack animals and vehicles drawn by other animals	89	46	0.52	2	0.02
3. boats and other water transports	31	18	0.58	0	0.00
4 porters and coolies	35	7	0.20	0	0.00
5. others	16	249	15.56	40	2.50
6. total	546	371	0.68	142	0.26
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	licences, taxes and rates		interest on loan capital	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	993	99.30	123	12.30
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	30	0.21	0	0.00
3. porters and coolies	117	4	0.03	0	0.00
4. animal-drawn carts and vehicles	88	52	0.59	12	0.14
5. pack animals	8	0	0.00	0	0.00
6. others	5	7	1.40	0	0.00
7. total	368	1086	2.95	135	0.37
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(9) : HOUSEHOLD AND HIRED MAN-DAYS PER WORKING DAY IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	household man-days per working day		hired man-days per working day	
		total (000)	per household	total (000)	per household
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	398	1.06	30	0.08
2. pack animals and vehicles drawn by other animals	89	81	0.91	1	0.01
3. boats and other water transports	31	36	1.16	33	1.06
4. porters and coolies	35	29	0.83	0	0.00
5. others	16	20	1.25	8	0.50
6. total	546	564	1.03	72	0.13
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	household man-days per working day		hired man-days per working day	
		total (000)	per household	total (000)	per household
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	11	1.10	21	2.10
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	136	0.97	13	0.09
3. porters and coolies	107	120	1.03	0	0.00
4. animal-drawn carts and vehicles	88	91	1.03	11	0.12
5. pack animals	8	13	1.62	1	0.12
6. others	5	6	1.20	5	1.00
7. total	368	377	1.02	51	0.14
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(10): MONTHLY HIRED LABOUR CHARGES IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	hired labour charges	
		total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)
1. bullock and buffalo carts	375	531	1.42
2. pack animals and vehicles drawn by other animals	89	44	0.49
3. boats and other water transports	31	1238	39.94
4. porters and coolies	35	0	0.00
5. others	16	523	32,69
6. total	546	2336	4.28
7. number of sample households	2742		
8. number of sample villages	959		

urban areas

type of transport	estimated number of households (000)	hired labour charges	
		total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)
1. motor cars and vehicles	10	1744	174.40
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	203	1.45
3. porters and coolies	117	0	0.00
4. animal-drawn carts and vehicles	88	384	4.36
5. pack animals	8	22	2.75
6. others	5	91	18.20
7. total	368	2444	6.64
8. number of sample households	1652		
9. number of sample blocks	443		

NATIONAL SAMPLE SURVEY : HOUSEHOLD TRANSPORT OPERATIONS

TABLE 1(11): MONTHLY GROSS EARNINGS FROM GOODS AND PASSENGER TRAFFIC IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	monthly gross earnings from			
		goods		passengers	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	8616	22.98	1081	2.88
2. pack animals and vehicles drawn by other animals	89	1692	19.01	2498	28.07
3. boats and other water transports	31	4662	150.39	217	7.00
4. porters and coolies	35	783	22.37	0	0.00
5. others	16	2572	160.75	1959	122.44
6. total	546	18325	33.56	5755	10.54
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	monthly gross earnings from			
		goods		passengers	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	5786	578.60	6201	620.10
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	3273	23.38	4997	35.69
3. porters and coolies	117	4040	34.53	0	0.00
4. animal-drawn carts and vehicles	88	4486	50.98	3046	34.61
5. pack animals	8	659	82.38	4	0.50
6. others	5	302	60.40	29	5.80
7. total	368	18546	50.39	14277	38.80
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 1(12): MONTHLY GROSS EARNINGS IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	number of working persons (000)	monthly gross earnings: total value (Rs. 000)	monthly gross earnings: per household (Rs.)	monthly gross earnings: per working person (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	375	425	9697	25.86	22.82
2. pack animals and vehicles drawn by other animals	89	96	4190	47.08	43.64
3. boats and other water transports	31	81	4879	157.39	60.23
4. porters and coolies	35	42	783	22.37	18.64
5. others	16	26	4531	283.19	174.27
6. total	546	670	24080	44.10	35.94
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	number of working persons (000)	monthly gross earnings: total value (Rs. 000)	monthly gross earnings: per household (Rs.)	monthly gross earnings: per working person (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	10	34	11987	1198.70	352.56
2. man-drawn transports like push carts, cycle-rickshaws etc.	140	154	8270	59.07	53.70
3. porters and coolies	117	128	4040	34.53	31.56
4. animal-drawn carts and vehicles	88	102	7532	85.59	73.84
5. pack animals	8	14	663	82.88	47.36
6. others	5	12	331	66.20	27.58
7. total	368	444	32823	89.19	73.93
8. number of sample households	1652				
9. number of sample blocks	443				

APPENDIX II

TABLE 2(1): NUMBER OF HOUSEHOLDS ENGAGED IN SMALL-SCALE TRANSPORT ENTERPRISE, HOUSEHOLD SIZE AND AVERAGE NUMBER OF PERSONS ENGAGED PER WORKING DAY : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	number of persons		number of persons per working day	
		total (000)	per household	total (000)	per household
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	5223	6.67	425	0.54
2. pack animals and vehicles drawn by other animals	131	762	5.82	96	0.73
3. boats and other water transports	42	258	6.14	81	1.93
4. porters and coolies	61	322	5.28	42	0.69
5. others	44	256	5.82	26	0.59
6. total	1061	6821	6.43	670	0.63
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	number of persons		number of persons per working day	
		total (000)	per household	total (000)	per household
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	95	7.92	34	2.83
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	533	3.70	154	1.07
3. porters and coolies	120	488	4.07	128	1.07
4. animal-drawn carts and vehicles	107	562	5.25	102	0.95
5. pack animals	10	52	5.20	14	1.40
6. others	6	31	5.17	12	2.00
7. total	399	1761	4.41	444	1.11
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(2) : VALUE OF LAND AND BUILDINGS USED IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	value of land		value of buildings	
		total (Rs. 000)	per household (Rs.)	total (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	3646	4.66	34371	43.90
2. pack animals and vehicles drawn by other animals	131	470	3.59	6581	50.24
3. boats and other water transports	42	4	0.10	72	1.71
4. porters and coolies	61	0	0.00	0	0.00
5. others	44	134	3.04	668	15.18
6. total	1061	4254	4.01	41692	39.30
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	value of land		value of buildings	
		total (Rs. 000)	per household (Rs.)	total (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	1152	96.00	3705	308.75
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	19	0.13	179	1.24
3. porters and coolies	120	0	0.00	0	0.00
4. animal-drawn carts and vehicles	107	1577	14.74	5112	47.78
5. pack animals	10	6	0.60	535	53.50
6. others	6	0	0.00	24	4.00
7. total	399	2754	6.90	9555	23.95
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(3) : VALUE OF DRAUGHT AND PACK ANIMALS AND VEHICLES AND TOOLS USED IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	value of animals		value of vehicles and tools	
		total (Rs. 000)	per household (Rs.)	total (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	259726	331.71	104932	134.01
2. pack animals and vehicles drawn by other animals	131	32296	246.53	5600	42.75
3. boats and other water transports	42	0	0.00	34254	815.57
4. porters and coolies	61	0	0.00	289	4.74
5. others	44	0	0.00	25944	589.64
6. total	1061	292022	275.23	171019	161.19
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	value of animals		value of vehicles and tools	
		total (Rs. 000)	per household (Rs.)	total (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	0	0.00	110913	9242.75
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	0	0.00	10512	73.00
3. porters and coolies	120	0	0.00	24	0.20
4. animal-drawn carts and vehicles	107	26724	249.76	18538	173.25
5. pack animals	10	4313	431.30	192	19.20
6. others	6	0	0.00	4897	816.17
7. total	399	31037	77.79	145076	363.60
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(4) : VALUE OF MONTHLY CONSUMPTION OF FUEL, LUBRICANTS ETC., OBTAINED BY PURCHASE AND BY HOME SUPPLY IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	purchase		home supply	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per househod (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	258	0.33	16	0.02
2. pack animals and vehicles drawn by other animals	131	29	0.22	0	0.00
3. boats and other water transports	42	249	5.93	135	3.21
4. porters and coolies	61	0	0.00	0	0.00
5. others	44	1190	27.04	4	0.09
6. total	1061	1726	1.63	155	0.15
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	purchase		home supply	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per househ (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	4398	366.50	0	0.00
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	53	0.37	1	0.01
3 porters and coolies	120	0	0.00	0	0.00
4. animal-drawn carts and vehicles	107	78	0.73	1	0.01
5. pack animals	10	0	0.00	0	0.00
6. others	6	4	0.67	0	0.00
7. total	399	4533	11.36	2	0.01
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(5): MONTHLY VALUE OF CATTLE FEED OBTAINED BY PURCHASE AND BY HOME SUPPLY IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	purchase total value (Rs. 000)	purchase per household (Rs.)	home supply total value (Rs. 000)	home supply per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	5442	6.95	8102	10.35
2. pack animals and vehicles drawn by other animals	131	2190	16.72	848	6.47
3. boats and other water transports	42	0	0.00	0	0.00
4. porters and coolies	61	0	0.00	0	0.00
5. others	44	0	0.00	0	0.00
6. total	1061	7632	7.19	8950	8.44
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	purchase total value (Rs. 000)	purchase per household (Rs.)	home supply total value (Rs. 000)	home supply per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	0	0.00	0	0.00
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	0	0.00	0	0.00
3. porters and coolies	120	0	0.00	0	0.00
4. animal-drawn carts and vehicles	107	4091	38.23	536	5.01
5. pack animals	10	320	32.00	0	0.00
6. others	6	0	0.00	0	0.00
7. total	399	4411	11.06	536	1.34
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(6): PERCENTAGE BREAKDOWN OF EXPENSES ON FUEL, LUBRICANTS ETC. AND CATTLE FEED OBTAINED BY PURCHASE AND BY HOME SUPPLY IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	percentage breakdown					
	fuel, lubricants etc.			cattle feed		
	purchase	home supply	total	purchase	home supply	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. bullock and buffalo carts	94.16	5.84	100.00	40.18	59.82	100.00
2. pack animals and vehicles drawn by other animals	100.00	0.00	100.00	72.09	27.91	100.00
3. boats and other water transports	64.84	35.16	100.00	—	—	—
4. porters and coolies	—	—	—	—	—	—
5. others	99.66	0.34	100.00	—	—	—
6. total	91.76	8.24	100.00	46.02	53.98	100.00
7. number of sample households	2742					
8. number of sample villages	959					

urban areas

type of transport	percentage breakdown					
	fuel, lubricants etc.			cattle feed		
	purchase	home supply	total	purchase	home supply	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. motor cars and vehicles	100.00	0.00	100.00	—	—	—
2. man-drawn transports like push carts, cycle-rickshaws etc.	98.15	1.85	100.00	—	—	—
3. porters and coolies	—	—	—	—	—	—
4. animal-drawn carts and vehicles	98.73	1.27	100.00	88.42	11.58	100.00
5. pack animals	—	—	—	100.00	0.00	100.00
6. others	100.00	0.00	100.00	—	—	—
7. total	99.96	0.04	100.00	89.16	10.84	100.00
8. number of sample households	1652					
9. number of sample blocks	443					

TABLE 2(7) : TOTAL VALUE OF MONTHLY CONSUMPTION OF FUEL, LUBRICANTS ETC., AND CATTLE FEED IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	fuel, lubricants etc.		cattle feed	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	274	0.35	13544	17.30
2. pack animals and vehicles drawn by other animals	131	29	0.22	3038	23.19
3. boats and other water transports	42	384	9.14	0	0.00
4. porters and coolies	61	0	0.00	0	0.00
5. others	44	1194	27.14	0	0.00
6. total	1061	1881	1.77	16582	15.62
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	fuel, lubricants etc.		cattle feed	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	4398	366.50	0	0.00
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	54	0.38	0	0.00
3. porters and coolies	120	0	0.00	0	0.00
4. animal-drawn carts and vehicles	107	79	0.74	4627	43.24
5. pack animals	10	0	0.00	320	32.00
6. others	6	4	0.67	0	0.00
7. total	399	4535	11.36	4947	12.40
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(8) : MONTHLY EXPENDITURE ON REPAIR AND MAINTENANCE AND RENT IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	repair and maintenance: total value (Rs. 000)	repair and maintenance: per household (Rs.)	rent: total value (Rs. 000)	rent: per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	6219	7.94	131	0.17
2. pack animals and vehicles drawn by other animals	131	1454	11.10	14	0.11
3. boats and other water transports	42	437	10.40	80	1.90
4. porters and coolies	61	1	0.02	0	0.00
5. others	44	1097	24.93	72	1.64
6. total	1061	9208	8.68	297	0.28
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	repair and maintenance: total value (Rs. 000)	repair and maintenance: per household (Rs.)	rent: total value (Rs. 000)	rent: per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	2048	170.67	294	24.50
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	377	2.62	2219	15.41
3. porters and coolies	120	0	0.00	8	0.07
4. animal-drawn carts and vehicles	107	1115	10.42	352	3.29
5. pack animals	10	123	12.30	2	0.20
6. others	6	41	6.83	1	0.17
7. total	399	3704	9.28	2876	7.21
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(9): MONTHLY EXPENDITURE ON STORES AND SERVICES IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	other stores		services	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	25	0.03	18	0.02
2. pack animals and vehicles drawn by other animals	131	3	0.02	19	0.14
3. boats and other water transports	42	2	0.05	5	0.12
4. porters and coolies	61	0	—	1	0.02
5. others	44	8	0.18	33	0.75
6. total	1061	38	0.04	76	0.07
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	other stores		services	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	157	13.08	146	12.17
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	1	0.01	9	0.06
3. porters and coolies	120	0	0.00	1	0.01
4. animal-drawn carts and vehicles	107	5	0.05	23	0.21
5. pack animals	10	10	1.00	0	0.00
6. others	6	0	0.00	3	0.50
7. total	399	173	0.43	182	0.46
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(10) : MONTHLY EXPENDITURE ON LICENCES, TAXES ETC., AND INTEREST ON LOAN CAPITAL IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953- MARCH 1954

rural areas

type of transport	estimated number of households (000)	licences, taxes and rates		interest on loan capital	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	151	0.19	100	0.13
2. pack animals and vehicles drawn by other animals etc.	131	46	0.35	42	0.32
3. boats and other water transports	42	18	0.43	0	0.00
4. porters and coolies	61	7	0.11	0	0.00
5. others	44	249	5.66	40	0.91
6. total	1061	471	0.44	182	0.17
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	licences, taxes and rates		interest on loan capital	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	1042	86.83	138	11.50
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	30	0.21	0	0.00
3. porters and coolies	120	4	0.03	0	0.00
4. animal-drawn carts and vehicles	107	61	0.57	30	0.28
5. pack animals	10	0	0.00	0	0.00
6. others	6	7	1.17	0	0.00
7. total	399	1144	2.87	168	0.42
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(11) : HOUSEHOLD AND HIRED MAN-DAYS PER WORKING DAY IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	household man-days per working day		hired man-days per working day	
		total (000)	per household	total (000)	per household
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	398	0.51	30	0.04
2. pack animals and vehicles drawn by other animals	131	81	0.62	1	0.01
3. boats and other water transports	42	36	0.86	33	0.78
4. porters and coolies	61	29	0.48	0	0.00
5. others	44	20	0.45	8	0.18
6. total	1061	564	0.53	72	0.07
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	household man-days per working day		hired man-days per working day	
		total (000)	per household	total (000)	per household
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	11	0.92	21	1.75
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	136	0.94	13	0.09
3. porters and coolies	120	120	1.00	0	0.00
4. animal-drawn carts and vehicles	107	91	0.85	11	0.10
5. pack animals	10	13	1.30	1	0.10
6. others	6	6	1.00	5	0.83
7. total	399	377	0.94	51	0.13
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(12) : MONTHLY HIRED LABOUR CHARGES IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953– MARCH 1954

rural areas

type of transport	estimated number of households (000)	hired labour charges	
		total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)
1. bullock and buffalo carts	783	531	0.68
2. pack animals and vehicles drawn by other animals	131	44	0.34
3. boats and other water transports	42	1238	29.48
4. porters and coolies	61	0	0.00
5. others	44	523	11.89
6. total	1061	2336	2.20
7. number of sample households	2742		
8. number of sample villages	959		

urban areas

type of transport	estimated number of households (000)	hired labour charges	
		total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)
1. motor-cars and vehicles	12	1744	145.33
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	203	1.41
3. porters and coolies	120	0	0.00
4. animal-drawn carts and vehicles	107	384	3.59
5. pack animals	10	22	2.20
6. others	6	91	15.17
7. total	399	2444	6.13
8. number of sample households	1652		
9. number of sample blocks	443		

TABLE 2(13) : MONTHLY GROSS EARNINGS FROM GOODS AND PASSENGER TRAFFIC IN HOUSEHOLD TRANSPORT ENTERPRISE : OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	gross earnings from			
		goods		passengers	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	8616	11.00	1081	1.38
2. pack animals and vehicles drawn by other animals	131	1692	12.91	2498	19.07
3. boats and other water transports	42	4662	111.00	217	5.17
4. porters and coolies	61	783	12.84	0	0.00
5. others	44	2572	58.45	1959	44.52
6. total	1061	18325	17.27	5755	5.42
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of households (000)	gross income from			
		goods		passengers	
		total value (Rs. 000)	per household (Rs.)	total value (Rs. 000)	per household (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	5786	482.17	6201	516.75
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	3273	22.73	4997	34.70
3. porters and coolies	120	4040	33.67	0	0.00
4. animal-drawn carts and vehicles	107	4486	41.92	3046	28.47
5. pack animals	10	659	65.90	4	0.40
6. others	6	302	50.33	29	4.83
7. total	399	18546	46.48	14277	35.78
8. number of sample households	1652				
9. number of sample blocks	443				

TABLE 2(14): MONTHLY GROSS EARNINGS IN HOUSEHOLD TRANSPORT ENTERPRISE: OCTOBER 1953–MARCH 1954

rural areas

type of transport	estimated number of households (000)	number of working persons (000)	gross earnings total value (Rs. 000)	gross earnings per household (Rs.)	gross earnings per working person (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. bullock and buffalo carts	783	425	9697	12.38	22.82
2. pack animals and vehicles drawn by other animals	131	96	4190	31.98	43.64
3. boats and other water transports	42	81	4879	116.17	60.23
4. porters and coolies	61	42	783	12.84	18.64
5. others	44	26	4531	102.97	174.27
6. total	1061	670	24080	22.69	35.94
7. number of sample households	2742				
8. number of sample villages	959				

urban areas

type of transport	estimated number of huseholds (0000)	number of working persons (000)	gross earnings total value (Rs. 000)	gross earnings per household (Rs.)	gross earnings per working person (Rs.)
(1)	(2)	(3)	(4)	(5)	(6)
1. motor cars and vehicles	12	34	11987	998.92	352.56
2. man-drawn transports like push carts, cycle-rickshaws etc.	144	154	8270	57.43	53.70
3. porters and coolies.	120	128	4040	33.67	31.56
4. animal-drawn carts and vehicles	107	102	7532	70.39	73.84
5. pack animals	10	14	663	66.30	47.36
6. others	6	12	31	55.16	27.58
7. total	399	444	32823	82.26	73.93
8. number of sample households	1652				
9. number of sample blocks	443				

APPENDIX III

COMPARISON OF RESULTS OF SUB-SAMPLES

1. In the seventh round of the survey the total sample was divided into two independent half-samples. In the rural sector, each of the two sample tehsils in any stratum constituted à half-sample giving an estimate for that stratum. In the urban sector likewise, the sample towns in each stratum were divided equally into two half-samples. However, when there was an odd number of sample towns in any stratum, the numbers in the two half-samples differed by one.

2. Table (1) gives the all-India estimates of some selected items by half-samples for the rural sector. Cols. (3) and (4) give the estimates half-sample-wise. Col. (5) gives the combined estimate and col. (6) the percentage difference.

TABLE (1) : HALF-SAMPLE ESTIMATES OF SOME SELECTED ITEMS IN RESPECT OF HOUSEHOLD TRANSPORT ENTERPRISE IN RURAL AREAS

item	unit	estimates of total			percentage difference
		half-sample I	half-sample II	combined	
(1)	(2)	(3)	(4)	(5)	(6)
1. number of households engaged in transport enterprise	lakh	11.10	10.11	10.61	9.33
2. number of persons per working day	,,	7.81	5.60	6.70	32.98
3. value of fixed capital	lakh (Rs.)	5511.39	4668.34	5089.87	16.56
4. monthly gross earnings	,,	275.74	205.87	240.80	29.02
5. number of sample households		1415	1327	2742	
6. number of sample villages		479	480	959	

3. The estimates by urban half-samples are presented in Table (2). The items are the same as in the case of rural sector.

TABLE (2) : HALF-SAMPLE ESTIMATES OF SOME SELECTED ITEMS IN RESPECT OF HOUSEHOLD TRANSPORT ENTERPRISE IN URBAN AREAS

item	unit	estimates of total			percentage difference
		half-sample I	half-sample II	combined	
(1)	(2)	(3)	(4)	(5)	(6)
1. number of households engaged in transport enterprise	lakh	4.17	3.90	3.99	6.77
2. number of persons per working day	,,	4.47	5.03	4.44	−12.61
3. value of fixed capital	lahk (Rs.)	1423.86	2517.60	1884.22	−58.05
4. monthly gross earnings	,,	264.59	384.12	328.23	−36.42
5. number of sample households		885	767	1652	
6. number of sample blocks		226	217	443	

4. It may be mentioned here that in the case of rural sector the combined estimates are built from the half-sample estimates and are the arithmetic means of them. In the case of urban sector, the combined estimates have been calculated directly from all the sample units together and not from the half-sample estimates.

5. The percentage differences given in column (6) of Tables (1) and (2) are the ratios of the excess of first half-sample estimate over that of the second to the combined estimate, expressed in percentages.

6. A scrutiny of the values of percentage difference shows that the estimates, in general, are of the same order. It varies from 9.33 to 32.98 in the rural sector while it ranges from 6.77 to 58.05 in the urban sector. The maximum percentage difference is in respect of average number of persons per working day in the rural sector and value of fixed capital in the urban sector. The minimum differences in both the sectors are for the number of households engaged in transport operations.

APPENDIX IV

FACSIMILE OF THE SCHEDULE OF INVESTIGATION

RURAL*
URBAN

ORIGINAL*
DUPLICATE

GOVERNMENT OF INDIA NATIONAL SAMPLE SURVEY : 1953-54

Household Schedule 2·3: Transport

(Seventh Round)

[1] identification: sample village			
1. zone-State natural division		9. district....................	
2. stratum		10. sub-division....................	
3. scheme		11. *tehsil/taluk/thana*....................	
4. sample unit I		12. police station....................	
5. sample unit II		13. village....................	
6. sub-sample		14. hamlet....................	
7. fraction code		15. census house no....................	
8. sample unit III			

[2] identification: sample block	
1. district....................	
2. town....................	
3. zone-State	
4. stratum	
5. sample town	
6. sample block	
7. sub-sample	
8. fraction code	
9. sample household	

[3] classificatory characters	
1. household size....................	
2. „ land possessed (acre 00.00)....	
3. „ industry-occupation code	
4. „ expenditure „	
5. „ religion „	
6. „ group „	
7. „ mother tongue „	

[4] informant's response and ability	
1. head of household....................	
2. informant's relation to head	
3. if head not available, why	
4. informant's willingness	
5. informant's ability	
6. if any second informant, why	
7. his relation to head	

[5] investigator: 1. name....................

2. roll no .... 3. signature....................

4. date (i) survey..............(ii)despatch of schedule....................

[6] inspector: 1. name....................

2. signature....................

5. date (i) inspection,........(ii) scrutiny..........(iii) despatch....................

[6] zonal scrutiny : 1. name of scrutiniser....................

2. signature.................... 3. date (i) receipt........(ii) scrutiny.......

(iii) despatch to I. S. I................4. no. of sheets of scrutiny notes attached.........

[7] general description and particulars		
1. name of transport		
2. name of establishment, if any		
3. year of foundation	4. no. of branches, if any	
5. proprietory-1/partnership-2	6. if 2, percent share............	
7. working season: whole year/month......	to........	
8. gross earnings, last year /season (Rs. 00.00)		
9. average no. of persons per working day..............during month ended on..........		

[8] land used in transport industry	
(1)	(2)
1. area (acres 00.00)	
2. purchase value (Rs. 00.00)	
3. rent paid during month ended on........(Rs. 00.00)	
4. improvement (Rs.00.00) during month ended on........	

[9] building used in transport industry

sr. no.	items	whole	part	total
(0.1)	(0.2)	(1)	(2)	(3)
1.	number			
2.	floor space (sq.ft. 00)			
3.	if purchased, value (Rs. 00.00)			
4.	if constructed, cost (Rs. 00.00)			
5.	expected years of life since construction or purchase (00)			
6.	rent paid during month ended on..............(Rs. 00.00)			
7.	repair and maintenance charges during month ended on. ....(Rs. 00.00)			

[10] volume of transport service produced during month ended on......

items	volume	miles
(1)	(2)	(3)
1. goods (std. mds. 00)		
(1)		
(2)		
(3)		
(4) sub-total		
2. passengers (no.)		
3. distance travelled empty	×	

[11] draught and pack animals

sr. no.	description of animals in transport industry	no. as on date of survey	value as on date of first use Rs.(00.00)		expected years of life since first use (00)	col (2) or col (3) ÷ col 4 (00.00)	during month ended on..............			
			purchased	homebred			rent paid (Rs. 00·00)	treatment and maintenance Rs. (00.00)	dead or discarded	
									no.	value (imputed) Rs (00.00
(0.1)	(0.2)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1.										
2.										
3.										
4.										
5.										
6.										
7.										
8.										
9.	total	×							×	

*delete whichever is inapplicable

12 transport vehicles, tools and furniture and fittings

		sr. no.	items	foreign-1, indigenous-2, mixed-3	no. as on date of survey	value at installation or first use Rs. (00.00)		expected year of life since installation or first use (00)	col. (3) or col. (4) ÷ col. (5) (00.00)	during month ended on			
						purchased	constructed			rent paid Rs. (00.00)	repair & maintenance Rs. (00.00)	broken or discarded	
												no.	imputed value Rs. (00.00)
		(0.1)	(0.2)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
vehicles, vessels & carts	power driven	1.											
		2.											
		3.											
		4.											
	animal driven	5.											
		6.											
		7.											
	man driven	8.											
		9.											
		10.											
tools, accessories, furniture & fittings		11.											
		12.											
		13.											
		14.											
		15.	total	×	×							×	

[13] consumption of fuel, lubricants and other materials

	sr. no.	items	specification of local and standard units			consumed during week ended on......						consumed during month ended on......					
			standard	local	ratio (00.00)	by purchase		by home supply		total		by purchase		by home supply		total	
						quantity (local) (00.00)	value Rs. (00.00)	quantity (local) (00.00)	value Rs. (00.00)	quantity (local) (00.00)	value Rs. (00.00)	quantity (local) (00.00)	value Rs. (00.00)	quantity (local) (00.00)	value Rs. (00.00)	quantity (local) (00.00)	value Rs. (00.00)
	(0.1)	(0.2)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
fuel for power, light, fan	1.	coal, coke	sr.														
	2.	charcoal, wood	sr.														
	3.	diesel & other oil	gallon														
	4.	electricity	s. units														
	5.	kerosene	sr.														
lubricants	6.																
	7.																
other materials	8.																
	9.																
	10.	sub-total	×	×	×	×		×		×		×		×		×	
cattle feed	11.	straw, fodder,	md.														
	12.	other feed	×	×	×	×		×		×		×		×		×	
	13.	total	×	×	×	×		×		×		×		×		×	

[14] other expenses (Rs.00.00) during month ended on......

sr. no.	items	amount Rs. (00.00)
(0.1)	(0.2)	(1)
1.	stationery articles	
2.	postage & stamps	
3.	other ores	
4.	sub-total-1	
5.	industrial servicing	
6.	insurance service	
7.	legal service	
8.	other services	
9.	sub-total-2	
10.	licences, taxes & rates	
11.	total : (items 4, 9, 10)	
12.	interest on loan capital	
13.	income tax	

[15] gross earnings during Rs. (00.00)

source	week	month
	ended on	
(1)	(2)	(3)
goods traffic		
1.		
2.		
3.		
4.		
5.		
6.		
7.		
8.		
9.		
10. sub-total-1		
11. passenger traffic		
12. total		

[16] personnel and labour charges during month ended on........

sr. no.	items		mandays per working day (00.00)		charges (hired) (Rs. 00.00)		
			hired	household	wages	other payments	total
(0.1)	(0.2)		(1)	(2)	(3)	(4)	(5)
1.	management and supervision						
2.	staff other than labour						
3.	skilled labour						
4.	unskilled labour	adult male					
5.		" female					
6.		child (under 14)					
7.		sub-total (items 4–6)					
3.	total (items1–6)						

MISCELLANEOUS

A BRIEF HISTORY OF THE ORGANISATION OF OFFICIAL STATISTICS IN INDIA DURING THE BRITISH PERIOD*

By S. SUBRAMANIAN

Department of Commercial Intelligence & Statistics, Government of India, Calcutta

SUMMARY. The history of Indian Statistics dates back to ancient times, but modern history may be said to have commenced from the early sixties of the nineteenth century with the setting up of a Statistical Committee in 1862 charged with the preparation of model statistical forms for administrative purposes. Then followed the appointment of the first Director General of Statistics in 1869 whose task was to produce a comprehensive statistical survey of the entire country. Later landmarks in the development of Indian Statistics were the Conference on Agricultural Statistics in 1884, the creation of a Statistical Bureau in Calcutta in 1895, its absorption by a Director General of Commercial Intelligence in 1905, the appointment of a Director of Statistics in 1914, the discontinuance as a measure of retrenchment, in 1923, of wage statistics among other series, the Reports of the Economic Enquiry Committee for 1925, the Royal Commissions on Agriculture and Labour of 1928 and 1930 respectively, the Bowley-Robertson Report of 1934, and the Inter-Department Committee on Official Statistics of 1945.

EARLY PERIOD

In India, as elsewhere, statistics really began as "data relating to matters of concern to the State and a collection of facts which might be of interest to the statesman. The *Artha-sāstra*[1] and *Ain-i-Akbari*[2] stand as towering examples of the truth of the above description and the temptation is great indeed to go to them for the unearthing of ancient reference to statistical compilation; but the research, although of great historical interest, would nonetheless be of little practical utility owing to the limited application and the inadequate coverage of such sources. This applies equally to the several *ad hoc* investigations carried out in the second half of the 18th century by official or private persons assisted from the public purse, e.g. Mr. Frankland's Statistical Report on the Pergunnahs southward of Calcutta (1758) which accumulated a great deal of isolated statistical material.

2. The position remained just the same till 1807 when the Court of Directors of the East India Company sent a recommendation to the Government at Fort William for a statistical survey of the country under the immediate authority of the Governor General in Council. Dr. Francis Buchanan was accordingly appointed along with an efficient staff of assistants for the work; and the points to be investigated were laid down in great detail. The instructions were well conceived, the remuneration was not inadequate and no time limit was imposed on the work but there were three fundamental defects in the scheme. First, there was no clear idea as to how much space could reasonably be allotted to each subject and to each district; secondly, no arrangement was made for utilising the material already in existence, or for obtaining from the regular Departments of Government their fair quota of aid; and lastly there was no arrangement for publishing the results of the survey, part by part, as they became available.

---

* This is the first of a series of articles which are proposed to be published on the development of official statistics in India. *Editor*.

[1] C. Third century B. C.

[2] Sixteenth century A. D.

3. Thus although Dr. Buchanan was at work for seven years and a sum of thirty thousand pounds was spent, he could leave no Statistical Account of any kind but only some trunks of manuscripts. The manuscripts were sent to England in 1816 and nothing was heard of these for twenty-two years. But in April 1938 the Court of Directors informed the Governor General in Council that Mr. Montgomery Martin had been permitted to inspect the manuscripts with a view to preparing them for publication. The result was the publication entitled *The History, Antiquities, and Statistics of Eastern India* (published in London) in three volumes containing an account of only nine districts of a single province, *viz.*, Bengal.

4. The various gazetteers published in the first few decades of the nineteenth century e.g., Thornton's and Hamilton's were equally unsatisfactory, for, these writers had to depend on chance notices by travellers and such other official accounts as they could find in the records of India House and it would obviously be useless to accuse the authors who had never visited the country and had no intimate knowledge of its conditions.

5. The only relieving features of this age which was otherwise wholly barren from the point of view of planned statistical development were the Annual Administration Reports of Local Governments based largely on the accounts furnished by District Officers. The Annual Reports afforded a standing opportunity for individual officers to enter into the details of the territories under their respective jurisdiction. The subjects embraced a very wide range and some of the reports formed complete statistical accounts of these territories. In addition to these accounts there were also special accounts of specified territories drawn up from time to time by Local Officers acting under the orders of their departmental chiefs. Examples of the latter type were Dr. Taylor's *Dacca* and Dr. John M. Cosh's *Assam* which were published between 1837 and 1840.[3] Unfortunately no permanent organization existed for working up the results obtained and for publishing them in the form of a complete book. In course of time the reports became scattered and unavailable.

6. In February 1855, the Court of Directors directed the Governor General to "procure by local enquiries, and with as little delay as possible, such information as may be effective for rendering the work a faithful register of the state of the country at present existing."[4]

7. In 1856, the Home Department of the Government of India came to an arrangement[5] with Mr. Townsend, Editor of the *Annals of Indian Administration*, a Journal which used to contain a consolidated summary of the information found in the Administration Reports of the various Provinces. Mr. Townsend's original proposal was that the Government of India should take 1,000 copies of the Annals but the Authorities of the Home Government were so convinced about the value of the publication that they raised the number to 1,500, although at the same time they reduced, somewhat the rate of payment. From that time the work was published entirely for Government and no attempt was made to sell to the general public. Mr. George Smith took over the editorship of the *Annals* after Mr. Townsend's departure in 1859 and as a result of three years' experience as editor he was led in 1862 to propose to Mr. Laing (then Financial Member of the

---

[3] Financial Department Progs., No. 79, October 1871.

[4] Despatch No. 1 of 1855 Statistical Department dated 21st February 1855.

[5] Home Department Progs. (Public) No. 86–87 dated 30th January 1856.

Council) the preparation of a series of Statistical Forms which would enable the writers of Annual Administration Reports to complete them on a uniform system adapted to that prevailing in England.

STATISTICS COMMITTEE OF 1862

8. In the circumstances, Mr. Laing who was obviously convinced of the soundness of Mr. Smith's suggestion, in a note written on 19th May 1862 pointed out among other things, that while he was engaged on a revision of the Tariff, he had to write to four or five different authorities to find out how many yards of piecegoods or lbs. of tobacco had been imported into India in a given period, and the returns were given in such different forms as to require protracted correspondence before the truth could be ascertained. And he added that for the purposes of both the Government and the commercial community it would be a great advantage to have a regular and methodical compilation and annual publication of such returns of Trade, Navigation and Finance as were published in England by the Board of Trade and the Treasury.[6]

9. Mr. Laing was also of opinion that the Financial Department would be the proper place for the establishment of a Statistical Branch for the purpose as they were already in the habit of compiling the different local returns into an Imperial Account for Budget purposes and as the information contained in such returns would be more frequently and more urgently wanted in the Financial than in any other Department.

10. He also recommended the appointment of a Committee for reporting to the Government on the preparation of Forms and Regulations with a view to the compilation of a uniform system of Imperial Statistics in the new Branch of the Financial Department.

11. Mr. Grey who was then in charge of the Home Department had other views in regard to the location of the Statistical Branch. He said that so far as a general statistical office had any existence at that time, it belonged to the Public Branch of the Home Department; and further the trade and navigation returns of the Provinces were received in the Revenue Branch of the Home Office. The Secretary to the Government in the Home Department was also Secretary in the Revenue Department although on the Revenue side he worked under the Financial Member; and so long as matters remained as they were, Mr. Grey said that the consolidation of such statistical information as returns of the trade and navigation, prices, etc., would more appropriately rest with the Home Department than with the Financial Department. According to Mr. Grey the most perfect arrangement would be the establishment of a separate Statistical Office which would communicate with the Government through the appropriate Branches of Departments.

12. Mr. Laing was prepared to admit that no one Department could compile the volume of Imperial Statistics altogether from its own resources but he was insistent that the volume could be most usefully and creditably compiled if it was done under his own eyes through the Financial Secretary rather than partly in one Department and partly in another.

13. The Governor General (Lord Elgin) wholly concurred with Mr. Laing and the following resolution was accordingly issued[7] :

---

[6] Financial Department Progs. No. 4, (Miscellaneous) dated June 6, 1862.

[7] Financial Department, Fort William the 6th June, 1862, No. 3814, Resolution.

"There is, His Excellency the Governor General observes, a mass of statistics in the Administration Reports of the different Local Governments and among the records of various Government Offices, but they are not compiled on any uniform plan nor brought together in any Central Office, so as to show the statistics of the Empire.

"It is obvious that it would be greatly for the advantage both of Government and of the commercial community to have a regular methodical compilation and annual publication of such Returns of Trade, Navigation and Finance as are published in England by the Board of Trade and the Treasury, and such compilation and publication would in the opinion of His Excellency the Governor General in Council be best effected in the Financial Department which would have most need for statistical information, and has large experience in compiling Statements and Returns received from all parts of India.

"The Financial Department must, however, His Excellency in Council observes, be careful to make no demand for statistical information except from its own Officers, without first ascertaining from the Department of Government concerned whether it can be obtained without overloading public Officers with extra work. As a general rule, the fewer the details required the more valuable and correct will be the information. An intelligent Officer could often himself give a concise report of the general features of trade in his district which would be accurate and valuable, while, if he were called upon for complicated returns he must leave them to careless subordinates, or neglect more important duties.

"The Returns of the English Board of Trade should, His Excellency in Council is of opinion, be taken as the basis of action, and each head gone through to see how far it is applicable to India without entailing any serious trouble or expense. Endeavours should be made to induce the different Local authorities to keep their statements in the uniform forms thus adopted, or, if kept differently for local convenience, to reduce their figures of measure or value to a common standard.

"The Chamber of Commerce will, His Excellency in Council believes, gladly give valuable assistance as to Commercial Statistics. Other branches of statistics of national importance might be gradually added as of population like those of the English Registrar General, and agricultural statistics, but this should only be done cautiously, and the Government will abstain from anything like merely scientific statistics confining itself to those which have an immediate bearing on practical Legislation and Government.

"A Statistical Society, aided and encouraged by the Government, but self-managed and voluntary, would be a most useful adjunct to any scheme of official statistics, and would embrace many fields of useful and interesting enquiry, which a Government could not undertake, while the advice and assistance of its leading members would be of great use in assisting the Government to frame its own returns and direct its own enquiries.

"Offers of aid have been received from some gentlemen who contemplate the formation of such a society, and as a first step towards the practical attainment of the above objects, His Excellency the Governor General in Council is pleased to appoint a Committee consisting of The Hon'ble Mr. Erskin, Mr. Bullen and Mr. G. Smith; The Hon'ble E. Drummond, Mr. Grote and Mr. E. C. Bayley as representing the Government (with power to add to their number) for the purpose of revising the Board of Trade and other Official Returns and of reporting to the Government as to the Forms and Regulations which they would recommend with a view to the compilation of a uniform system of Imperial Statistics in the newly created Branch of the Financial Department."

14. The Committee submitted its first Report in March 1863 and with the Report forwarded a number of draft forms (prepared on the plan recommended by the International Statistical Congress held in London in 1860 under the Presidency of the Prince Consort) for the compilation of financial and commercial statistics and of statistics pertaining to "Civil Government" excluding judicial and educational statistics which were to be dealt with separately in a second report. The Committee did not touch the subject of Vital Statistics, as it was informed that proposals on that subject were under submission by the Inspector General of the Medical Department. The Committee, in conclusion pointed out how greatly the value of all statistical information would be enhanced by accurate population figures but it was hesitant to offer suggestions in regard to a General Census, as it was not sure that it had been asked to do so and since it imagined that this question would, in any case, be left to the discretion of Local Governments.

15. In the meantime, Sir Charles Trevelyan (who was to take over later as Financial Member) had brought with him certain alternative sets of forms[8] prepared in the India Office and these forms although considered generally inferior to those prepared by the Committee, had one or two advantages in regard to arrangement of matter and moreover they included judicial and educational headings also. These latter forms were therefore passed on to the Committee for consideration with a view to a possible recasting of the Tables which it had originally proposed for Administration Reports. The Committee in its letter dated 17th August 1864 acknowledged receipt of the alternative forms and promised a further report on these forms at a later stage. It submitted a third report on 14th September 1865. With this report which contained the original forms as transposed and amended the Committee brought to a close the two specific duties assigned to it in the Government Resolution dated 6th June 1862.

16. The project of forming a separate Statistical Society which was alluded to in the Resolution had not been carried out. But the Asiatic Society had, on the advice of the President of the Committee, reconstituted the Statistical Committee of its Council and the Agri-Horticultural Society had offered its co-operation in obtaining agricultural statistics.

17. The Committee also stated that as it consisted of members whose hands were already full of absorbing business, it was not likely that statistical enquiries could be properly organized under their supervision. They felt that under their constitution they could give but occasional attention to such duties; whereas the task of obtaining and shaping reliable Imperial statistics for a great country like India should belong to a special and properly organized Department. This would not necessarily be a costly Department but its work, to be really useful, should be steadily and continuously prosecuted; and the Department should have a corresponding Agency in each Presidency and Chief Commissionership.

18. In his letter dated the 2nd April, 1866 addressed to the Statistical Committee, Mr. Bayley, Home Secretary, stated that His Excellency in Council was very sensible of the value which belonged to the possession by Government and the public of the full and extensive information embraced by the forms submitted by the Committee and that he was well disposed to attempt the collection of such information in the uniform shape proposed by them. It was obvious, however, that it would not be possible at once to collect

---

[8] Home Department Progs. A (Public) Nos. 115-36 dated October 31, 1866.

throughout India with any approach to accuracy, the whole of the information for which the forms provided and as it was most desirable to avoid the publication of erroneous statistics, the Governor General-in-Council would be glad to receive the opinion of the Committee as to the practical manner in which these forms might be introduced and gradually worked up.

19. Mr. Bayley also enclosed the views of Local Governments on the forms originally submitted by the Committee in 1863. Opinion was generally in favour of adoption of these forms although the Lieutenant Governor of the Punjab while reluctantly agreeing to the proposal was nevertheless inclined to strike a discordant note. His Government took nearly thirty months for considering the question and finally in their letter dated 3rd January, 1866, which will bear quotation said :

"Being of opinion that the pressure on our Officers of almost all the grades, resulting from the larger number of Returns and Reports they have to render, is already such as materially to impair the ability and inclination of many to devote to their executive duties the time and attention which the complete fulfilment of these involves. His Honour has felt exceedingly doubtful whether the state of the Administration is in the Province such as to justify the introduction of additional Returns of the complex character contemplated by the Statistical Committee, the machinery for collecting the information being non-existent. Hence His Honour has felt very reluctant to approach the subject, while fully admitting the great value of such statistics if they could be furnished with fair correctness, and without sacrificing other subjects even more important.

"His Honour believes that Returns which need special enquiries will never be prepared, either economically, effectively or without serious harassment to the people until we associate with us, for the purpose, representatives from among the people themselves, and there is, perhaps, no matter with which they may so readily and beneficially be thus associated."

20. Another very important question had already been raised by the Government of Bombay. The latter, in their letter dated 30th October, 1863, stated that the desired Returns could not be prepared without organizing a special establishment for the purpose and since it was impossible to say, until some experience had been obtained in the matter, what the size of the establishment would be, they requested the Government of India to give them the necessary authority to set up such Establishment as might be found necessary. The Government of Bombay sent a reminder on the subject on 22nd February, 1864 and the reply of the Financial Secretary of the Government of India dated 10th May, 1864 constituted an important statement of policy which must have acted as a damper on the enthusiasm of Local Governments and greatly retarded the development of Provincial Statistical Organizations. This was a most unfortunate setback to the cause of statistics in India. The Financial Secretary, Mr. Lushington declared in his letter :

"The bulk of the statistics proposed by the Statistical Committee can be prepared without the organization, at a heavy expense, of a Statistical Department under each Local Government.

"It is true that statistics whether pertaining to Finance or to Civil Administration must be put together in a Central Office but this would not involve more trouble than the preparation of the present Administration Reports.

"There are, no doubt other heads of information such as population and vital statistics, the industrial occupations, agriculture and internal trade for which tables have been proposed by the Committee, but in respect of these, it is obviously desirable that the agency for collecting in a reliable form should be stated or, if expedient, be organized with the previous sanction of the Government of India before any step is taken to form a statistical office of the Presidency for the record of the statistics."

21. The Committee's reply to the Government of India, resubmitting the forms as amended in accordance with the suggestions of Local Governments was signed by only four (including Mr. Grote, the Chairman) out of the expanded strength of ten, and was clearly unenthusiastic. The Committee at one time intended that apart from annual statistics of trade, finance, education etc., the returns of population, agriculture and live-stock, mines and manufactures should be made quinquennial or decennial but, on reconsideration came to the conclusion that the attempt to prescribe any such arrangement would be to interfere with the entire discretion of the local authorities which they would leave, for the time, unembarrassed. It was not proposed to ask for the preparation of any returns other than those which the Local Governments found practicable, safe, and convenient to procure. Again, information that could not be collected annually was not expected to be collected more frequently than was deemed expedient. The Committee would not, for a moment, suggest that statistical objects should clash with those which involved the peace and well-being of the country.

22. The Committee did not doubt that with the aid of the establishments possessed by District Officers in most parts of the country—if those Officers were but sufficiently impressed with the importance of the matter—the method of rendering the Returns would greatly improve by practice. "Everything must have a beginning; the rudest statistics, if honestly rendered, are better than no statistics; and the mere fact of publication and comparison will bring to light errors and blots which, in the darkness of Offices, would never have been discovered."

23. In conclusion, the Committee recommended the following course of action :

(a) The Local Administration should be requested to adapt the statistics already possessed by them, or available to them, to the forms prescribed for uniformity; and to fill up these forms each year as far as possible (either in part or in whole) and embody them in their Annual Administration Reports distinguishing accurate from approximate figures.

(b) They should, further, be requested to send to the Government of India, statements showing those additional forms of the prescribed list which might be filled up by a small rearrangement of machinery at some small expense, and without political danger or serious inconvenience; and to explain how far they proposed to take measures to this end.

(c) That each Administration should submit a further list of the forms of the prescribed series which it was not considered possible to fill up within any reasonable time without expensive new machinery, great expense or political inconvenience; and they should state whether they considered any of these statistics so important as to justify the Government in undertaking their preparation.

24. The Government of India would be in a position, on a comparison of the reports of the Local Governments, to come to a conclusion as to the extent to which a system of

statistical returns could be definitely prescribed and required from all Provinces. Information which could be supplied from most of the Provinces might be deemed worth collecting in the remaining few at a cost which would be justified by its Imperial and general importance.

25. The forms thus finalized were sent to the Local Governments with covering letters from the Home Department dated 31st October, 1866 containing a request that an attempt might be made to collect the information embraced by the forms in the uniform shape proposed by the Committee and, if it appeared feasible at once, to instruct the Officers to set on foot the system of returns and of record which would be necessary for the purpose.

26. The matter was in the ordinary course referred to the Secretary of State for India and the Rt. Hon'ble Sir Stafford Northcote, the then Secretary, in his despatch dated 23rd March, 1867[9] conveyed his general approval to the forms as finally proposed by the Committee and said that he would not object to the small additional expenditure that might have to be incurred on this account. He also suggested that the forms, instead of being incorporated into the Administration Reports, should be placed together in an appendix, so that the returns for a series of years could be bound in one volume, if desired. Lastly, he desired to impress on the Government of India the need for early transmission of the Reports since their value depended upon their being forwarded soon after the close of the period to which they referred.

27. The very next year, a Statistical Abstract relating to British India (1840–1865) was brought out by the India Office for presentation to the Houses of Parliament. (This blue-book became an annual feature and it continued to be published in England till 1923). Already, in July 1866 the Statistical Department of the India Office had brought out the first issue of the Statement of the Moral and Material Progress of India giving a connected narrative of the events of chief interest in the various branches of Administration based on the detailed reports of the Local Governments in India.

28. The account of the growth of Statistical Organisation should now be switched back to March 1852 when the Madras Government issued orders for a series of District Manuals which were intended to show the entire Presidency in minute detail. Subsequently in 1866 Sir Richard Temple inaugurated the statistical operations which later produced the celebrated Gazetteer of the Central Provinces. This latter work created a most excellent impression and on the strength of this the Government of India submitted to the Secretary of State in May 1867 a proposal for the compilation of similar Gazetteers for the remaining Provinces.[10] The proposal duly received the approval of the Home Government and in their letter dated 19th October, 1867 on the subject, the Government of India requested the Local Governments to take the necessary steps for the compilation of a Gazetteer as proposed. Various schemes were set on foot in this connection by the Local Governments but apart from their excessive costliness they wholly lacked a unity of plan and central supervision. The need for systematic organisation was also pressed on Government from various quarters particularly the Asiatic Society of Bengal.

---

[9] Home Department Proceedings (Public), July, No. 66.

[10] Department of Agriculture, Revenue and Commerce Progs. No. 112, dated 8th September 1871.

Appointment of Dr. Hunter as Director General of Statistics

29. The Governor General-in-Council agreed that immediate steps ought to be taken for ensuring some uniformity of plan in the Gazetteers of the various Provinces and in July 1869 Mr. W. W. Hunter (later Sir) of the Civil Service was entrusted with the work. Mr. Hunter (who was already in charge of the preparation of the Bengal Gazetteer) was directed in the first place, to ascertain exactly what had been done under the orders of the several Local Governments and Administrations, and then submit to the Home Department of the Government of India a comprehensive scheme for utilising the information already collected and for prescribing the principles to be followed in the future preparation of local Gazetteers and for consolidation of the entire work into a general Gazetteer of India.

30. In November 1869 Mr. Hunter submitted his plan for an Imperial Gazetteer of India. Although final consideration of his plan was postponed till the contemplated formation of the Department of Agriculture, Revenue and Commerce, his services were retained for the purpose of arranging for uniformity in the Provincial compilations at that stage which clearly demanded a Central guiding Authority.

31. In their Resolution dated 8th September 1871, The Department of Agriculture, Revenue and Commerce of the Government of India (of which Mr. A. O. Hume was the Secretary), appointed Mr. Hunter, permanent Director General of Statistics (in addition to his duties as Editor of the Bengal Gazetteer) for the purpose of supervising the operation connected with the Provincial Gazetteers and for consolidating these later into an Imperial Gazetteer of India. Mr. Hunter was given a salary of Rs. 1,500 per month and a small office establishment consisting of two assistants.

32. The task before Mr. Hunter consisted of two stages, *viz.*, first a local enquiry, conducted on a uniform plan in each of the 140 districts of British India and second, the consolidation of the material obtained. He had to secure the co-operation of Provincial Governments by respecting their individuality and by modifying the uniform scheme to suit their peculiar circumstances and at the same time collect the material systematically and cheaply by enlisting the free aid of the District Officers.

33. Mr. Hunter circulated to Provincial Governments a questionnaire entitled 'Heads of information required for the Imperial Gazetteer' under six categories, *viz.*, topographical, ethnical, agricultural, industrial, administrative and medical, in each District. A paid Provincial compiler was appointed at each Provincial Headquarters responsible for obtaining the returns from the District Officers (who did the work honorarily) and for supplementing this information by contacting the Heads of Departments and others who could help. The Provincial compiler was ultimately to consolidate the results into a Statistical Account for this Province.

34. The work went on for twelve years and the District and Provincial Accounts formed about 100 volumes consisting of 36,000 pages. The 'Native States' were, strictly, speaking, outside the scope of the survey but since no Account of the country could be complete without some reference to these territories, an attempt of some kind was made to co-ordinate the already available material in regard to these. The problem was easy of solution in the case of States which were politically dependent on Provincial Governments as the respective Provincial compilers easily obtained the necessary material from them

but in the case of other States, there were certain political difficulties which made it impossible to undertake any such enquiries. Thus the Accounts of the Native States in the series, though marking an advance on the then existing state of affairs could not, however, claim to equal the Accounts of the British Indian Provinces.

35. The hundred volumes of the Statistical Survey were meant primarily for the use of the Indian Administrators and for the Controlling Body in England and were beyond the reach of the public. To rectify this deficiency, Hunter set out to reduce the voluminous records of the Statistical Survey to a practicable size and accordingly in 1881 the first edition of the Imperial Gazetteer of India was brought out in nine volumes. Government had decided that the publication of the Gazetteer should not be delayed for the purpose of incorporating the results of the 1881 Census. The Census of 1872 did not supply a simultaneous count for the whole of India and it was found impracticable to bring the economic statistics in the Gazetteer up-to-date with the help of the Administration Reports. The result was that the latest figures under the different heads were based mostly on material collected at widely varying points of time between 1868, the year of the Punjab Census and 1887, although in one or two cases, e.g., trade statistics, it was found possible to take the facts down to 1879 and 1880. In the circumstances Hunter had to defer to a later edition the reduction of the statistics to a uniform basis. The difficulties had better be put in Hunter's own words.

"The areas, distances and similar details were taken from the latest scientific measurements; but the frequent changes in the jurisdiction of Indian Districts introduce an element of variation difficult to eliminate. In some cases, averages will not work out correctly, as in the rates of local taxation per head, where the municipal area often differs from the town area taken for the Census. In other instances, the items will not add up; as in certain Districts where the religious classification of the people does not yield the exact total arrived at by the general enumeration. There are not very many discrepancies of this sort, and no labour has been spared to get rid of them. But in several cases I cannot go beyond the figures supplied to me by the local authorities and although I may see that there is something wrong, I am unable to set it right. I beg that those who may come after me will, in improving my work, remember the conditions under which it has been done. When it was begun, no one knew exactly the population of a single Province of India, or of a single District of Bengal.

"I believe that in spite of all its defects, this work will prove a memorable episode in the long battle against ignorance; a breakwater against the tide of prejudice and false opinions flowing down upon us from the past; and a foundation for a truer and wider knowledge of India in time to come.

"I leave the work to the charitable judgment of those who can contrast it with the efforts of Indian statisticians who have gone before me; I also leave it with a sure confidence that it will be improved by brethren of my Service who come after me."

36. A second edition of the Imperial Gazetteer augmented to fourteen volumes was published in the year 1885-87. This edition took as its starting point the Census of 1881 which was the first complete and synchronous enumeration in India; the other statistics ranged between the years 1882 and 1885. With the completion of his writing work, Hunter took the first steps towards breaking up his office establishment in July 1886 and himself

retired a year later. With his retirement the post of Director General of Statistics was allowed to lapse and it was only in 1895 that the second Director General of Statistics with the Government of India was appointed.

37. A revised form of the Article on India, greatly enlarged and with statistics brought up to date, appeared as an independent volume in 1893 under the title '*The Indian Empire : Its Peoples, History and Products*.' It is said that Hunter wrote the greater part of the above book in his own hand.

38. The second edition of the Imperial Gazetteer having run out of stock, the Government of India decided in 1901 that a new edition should be brought out. Many changes were introduced in the general scheme as well as in the methods of compilation. The number of volumes was raised to twenty-six, the single volume of '*Indian Empire*' alone being expanded to four volumes called Descriptive, Historical, Economic and Administrative respectively.

39. The work however, was entrusted this time not to the Director General of Statistics (Mr. O'Conor) but to Sir Herbert Risley, Census Commission for 1901 and later when he was promoted as Home Secretary was succeeded as Editor by Mr. W. S. Meyer. The responsibility for the final form of the work and the duty of seeing it through the press rested on the Editor in England, Mr. J. S. Cotton (who had been closely associated with Hunter in the earlier editions) and a Committee appointed by the Secretary of State. The principle adopted in the preparation of the new edition was to entrust each subject to an author best qualified to deal with it by special study or by Official experience. As in previous editions, the basic material collected by District Officers and Officials of Indian States and supplemented by special contributions from experts and every page of written matter was submitted to the Administration or Department concerned for prior approval. The volumes as approved came out in 1907.

40. While all this was happening in the Home Department of the Government of India, several important developments were taking place in other Departments as well. The Department of Revenue, Agriculture and Commerce issued in 1876 the Miscellaneous Statistics relating to British India containing figures for 1874–75. The Finance and Commerce Department also began publishing an annual Finance and Revenue Accounts from the year 1880. In the preceding year the Home Department had started publication of their statistics under the long title Statistics of British India for the Judicial and Administrative Departments subordinate to the Home Office which included educational and public health statistics also.

41. The Commission which enquired into the causes of the Madras Famine of 1877 had suggested in its Report (submitted in 1880) the formation of a separate Department[11] whose duties should consist in the collection of facts relating to "the condition of the agricultural community and the agricultural produce of the country." The efficiency of such a special Department depended on the completeness and accuracy with which the agricultural, vital, and economic statistics concerning the Department were collected in each village and compiled in each subdivision and district throughout the country. In March 1881, armed with the recommendation of the Famine Commission, the Government of India

---

[11] Finance and Commerce Department, A Progs. 318–19, salaries, September 1892.

requested the sanction of the Secretary of State for the creation of the Agricultural Department. He gave the sanction asked for but drew attention to the fact that the Government of India had made no reference to the Commission's recommendation in regard to a more complete and systematic rendering of agricultural and economic statistics.

42. It was acknowledged that practically no agricultural statistics were maintained in any Department of the Government of India and that it would become the duty of the Department of Agriculture to inaugurate them. It was felt, at the same time, that it would be best to make such compilations the duty of officials who would be in contact with practical agriculture, and in December 1881, the Government of India issued a Resolution recommending the setting up of an Agriculture Department in each Province but somehow the importance of the work of collection and collation of agricultural statistics did not receive prominent mention in the Resolution.

(It must now be recalled that in 1866 certain tables in forms proposed by the Statistical Committee were recommended by the Government of India for adoption as appendices to the Annual Administration Reports of Local Governments. These forms in so far as they related to agricultural statistics were re-examined by a conference of Officers of Agricultural Departments held in Calcutta in the year 1883-84 and the modified returns—the chief modification being the complete omission of the table on rent rates—comprising eight tables were compiled for the first time in 1884-85 and were published in 1886 under the title "Returns of Agricultural Statistics of British India." These tables were continued on an annual basis till 1887-88 when a slight rearrangement was made in them; the tables were again revised in 1891).

43. However, an enormous mass of agricultural statistics began to pour into the Statistical Branch of the Revenue and Agricultural Department every year but owing to the absence of local checks, the task of scrutinising the returns fell entirely on a meagre establishment of two low paid clerks.

The Superintendent of the Agricultural and Statistical Branch had a host of other subjects under his charge ranging from horse breeding to emigration. None of the Secretariat Officers could deal adequately with the questions that came up from the Statistical Branch and the result was that the Department had to go on dealing with these important data in a rather perfunctory manner. With a view to rectifying this position, the Revenue and Agricultural Department submitted to the Finance and Commerce Department in May 1890 the following proposals for placing their statistical work on a proper basis :

(a) The Assistant Secretary in the latter Department (with his establishment) be transferred to the Revenue and Agricultural Department. Or

(b) The statistical work in question be taken over by the Finance and Commerce Department. Or

(c) A special establishment consisting of an Assistant Secretary, six clerks and ten tabulators costing Rs. 1,835 per month be created in the Revenue and Agricultural Department for the Statistical work.

44. In reply the Finance and Commerce Department stated that they preferred the second proposal and they added that they could take up the extra work with the addition of but nine hands to their Statistical Branch located permanently in Calcutta which

would be accompanied by a reduction in the staff of the other Department. This was not acceptable to the Revenue and Agricultural Department who had gone back on their original proposal and now maintained that they were responsible for the greater part of the statistics then collected by the Government of India and that they must continue to retain the direction of a branch of administration for which they were directly responsible. According to them the Finance Department's refusal of their application for a Statistical Officer and establishment was based "on the erroneous supposition that for agricultural statistics all that was required is their bare compilation." They considered it desirable to undertake a thorough overhaul by independent officials of the statistical work which really fell to the Department. Accordingly Mr. Baines, Census Commissioner for 1891 and Mr. O'Conor, Assistant Secretary in charge of the Statistical Branch of the Finance and Commerce Department were requested to undertake the investigation. They conducted their enquiries in an independent capacity in addition to their other work and they submitted their note in October 1891. In the note, they gave a brief description of the available information and the sources from which it was derived and tried to assess the statistical bearing and value of the general returns. In so far as agricultural statistics were concerned it was found that little use was made of the information after it had been consolidated into Imperial returns. There was no systematic collation of subjects so intimately connected with each other as rainfall, prices, area under crops, and the movement of these crops and, above all, there was no attempt to co-ordinate the statistics of sea-borne traffic with those of rail or road and river-borne traffic. This was, perhaps, the most significant instance of waste of good materials and of efficient office power.

The task of receiving the rail and river-borne trade figures after the introduction of the new system of registration (1888-89) fell upon an officer who was already responsible for the administration of two other large branches of the Revenue and Agricultural Department; the statistics of foreign trade were on the other hand, the responsibility of a special Branch of the Finance and Commerce Department. There were other cases in which the absence of co-operation resulted in superfluous work; for example, minerals like coal and salt were dealt with as economic products in the Department of Revenue and Agriculture while the same information was also collected by the Finance and Commerce Department.

45. In conclusion Messrs Baines and O'Conor pointed out that it was clearly beyond the capacity of a mere branch of a large Department to deal satisfactorily with the wide variety of returns that were being received from the different parts of a vast country like India. Again, the task of finding out in what way the different series of returns supplemented one another and of applying the knowledge thus gained in fact needed a full-time specialist, who would be responsible for the direction of a competent establishment, the scrutiny of the materials supplied, and the publication and review of the materials in the manner best calculated to convey to others, whether Government or the public, the information thus collected as fully, clearly and "palatably" as possible. Although it was impossible to lay down any rule in advance concerning the composition and strength of the staff required for the work, it was obvious that there should be two divisions, the first consisting of book-keepers and computors and the second, a small experienced body of supervisors on proportionately higher pay.

46. On the strength of the Baines-O'Conor Report Sir Philip Hutchins, then Home Member (also in charge of the Revenue and Agricultural Department) made a proposal in

May 1892 to the effect that a Statistical Bureau should be established under the administrative control of the Revenue and Agricultural Department and that Mr. O'Conor should be appointed Statistical Reporter to the Government of India. Mr. O'Conor was to be relieved of all routine and administrative work and placed entirely in charge of statistics. Sir Philip was not, however, agreeable to the suggestion that with Mr. O'Conor's transfer to the Revenue and Agricultural Department, the Home Member should take over the entire commerce work under his charge but he was willing to allow Finance and Commerce Department to consult Mr. O'Conor on any "commercial matters of importance" even after transfer.

47. The Finance Department agreed that it would be an excellent thing to have a separate Statistical Reporter with a separate office but owing to the deterioration of the financial position they ruled that the proposal should be dropped for the time being.

A Statistical Bureau under a Director General of Statistics

48. Things went on as before in both the Departments until in February 1895 the Government of India submitted a proposal to the Secretary of State for placing Indian agricultural statistics under the Assistant Secretary of the Finance Department and designating this Officer (Mr. O'Conor) as Director General of Statistics.[12] The arrangement was to be made without any increase in expenditure. The Secretary of State in his despatch dated 14th March, 1895 sanctioning the proposal, observed that it would be a decided advantage to have all the Indian statistics placed under the direction of a "single selected officer" and hoped that the statistics available in the Home Department would also be brought under this officer in course of time. Accordingly the following resolution was issued by the Government of India :

[13]"One of the most important of the functions assigned to the Department of Revenue and Agriculture in the despatch from the Secretary of State, No. 55, Revenue, dated 16th June, 1881 was the more complete and systematic ascertaining and rendering available of the statistics of vital agricultural and economic facts for every part of India, in order that Government and its Officers may always be in possession of an adequate knowledge of the natural condition of the country. During the last fourteen years the land-record system has been steadily extended and improved, and other subjects, not directly connected with the land, have received attention; so that the body of statistics connected with matters under the control of the Department of Revenue and Agriculture which annually reaches the Government of India and through them the Secretary of State, is now fairly complete for most of the Provinces of India. It has, however, long been felt that the disconnected manner in which these statistics are at present compiled and published seriously detracts from their value, and that their intelligent examination, collation, and discussion are necessary in order to secure uniformity and to render them as useful as they might and should be.

"It has therefore been decided, with the approval of the Secretary of State, to place the statistics with which the Department of Revenue and Agriculture is concerned under the control of an Officer skilled in such matters, who will deal with them continuously upon a uniform plan, and will collate them and publish the results in a properly digested form. The Department of Finance and Commerce already so deals in its Statistical Branch with a large body of periodical returns relating to matters, many of which are more or less intimately connected with the material collected by the Department of Revenue and Agriculture. This branch will now be strengthened and expanded by the transfer to it of the statistical establishment hitherto entertained by the Department of Revenue and Agriculture, and will then form a statistical bureau which will deal with the material collected by the two Departments of Finance and Commerce, and Revenue and Agriculture, and eventually, it is hoped, with that collected by the Home Department also.

---

[12] Department of Revenue and Agriculture, A Progs. Nos. 3 to 9 (General), June 1895, File No. 25 of 1895.

[13] Revenue and Agriculture Department Resolution No. 1552-6 dt. May 23, 1895.

"The new bureau will be presided over by an officer under the title of Director General of Statistics. In this capacity he will be, not a member of the Secretariat staff, but an independent officer who will publish his compilations and reviews in his own name and on his own responsibility; though he will deal with the statistics belonging to each Department in communication with and under the general direction and control of that Department. By the notification mentioned in the preamble, Mr. J. E. O'Conor, C.I.E., Assistant Secretary to the Government of India in the Department of Finance and Commerce, has been appointed to be the first (sic) Director General of Statistics. He is an Officer of the Enrolled List of that Department and will retain his present appointment ex-officio : he and his establishment will therefore be officially controlled by that Department.

"In future the following returns should be forwarded to the Director General of Statistics instead of the Department of Revenue and Agriculture :

(1) Annual returns of agricultural statistics.

(2) Annual returns of inland trade.

(3) Quarterly returns of inland trade.

(4) Quinquennial returns of crop outturns.

(5) Monthly statements of imports of wheat, cotton, linseed and indigo into the port towns of Calcutta, Bombay and Karachi.

(6) Forecasts of cultivation and of outturn of crops.

(7) Returns showing the state of cultivation of tea, coffee and cinchona; and

(8) All other tabular statements that are now furnished periodically to the Department of Revenue and Agriculture, such as statements of the financial results of irrigation operations, statements of irrigation operations in the several provinces, return of canal traffic and statements of additions to the open mileage of railways.

"The weekly weather and crop reports will continue to be sent to the Department of Revenue and Agriculture as heretofore. But from the date of receipt of this Resolution the other returns specified above should be addressed to the Director General of Statistics, and all correspondence regarding them should be conducted with him direct. It was at the same time laid down that all important communications to the Provincial Governments shall be issued from the appropriate Department of the Government of India."

49. With effect from 1st June, 1895 the Office of the Director General of Statistics was continued by the transfer of 1 Superintendent and 26 clerks from Finance Department and 5 clerks from the Department of Revenue and Agriculture and the creation of an additional staff of six clerks. This was the real beginning of the present Department of Com-

mercial Intelligence and Statistics which has experienced many vicissitudes in its career of fifty years as will be seen in later paragraphs.

50. Mr. O'Conor was already responsible in the Finance and Commerce Department for bringing out the annual Financial and Commercial Statistics of British India which replaced the Statistical Tables Relating to British India and Finance and Revenue Accounts —Part III. Again, as already mentioned, his Branch consolidated and published the monthly returns of sea-borne trade and shipping received from the Customs Authorities. An annual review based on the aggregate of the monthly returns was also issued by Mr. O'Conor. In addition to this, Mr. O'Conor's Branch was responsible for the following statistics :

(a) *Prices :* Retail prices of articles were obtained from 200 district centres and published every fortnight. From this statement an annual return was prepared giving figures for various subdivisions of the country based on meteorological considerations. Food-grains at selected centres were shown separately.

Half-yearly wholesale prices for the main articles of commerce at Calcutta and Bombay were compiled from the price lists published by the Chambers of Commerce. Special monthly returns of tea sales and salt revenue were issued in addition to the fortnightly report.

(b) *Wages :* In regard to agricultural workers and artisans, the returns received from districts through the Local Governments were summarised in the Department. Industrial wage statistics were ascertained through the courtesy of managers. The returns were published half-yearly and from 1891 were added in annual review.

(c) *Industries :* A review of the working of the Indian Companies Act based on the annual returns furnished statutorily by companies to Local Government was published by the Department. Another annual review was also published concerning the progress of the more important industries like cotton, jute, woollen and paper mills, breweries.

51. In view of the enormous amount of work that had now been thrown on him, Mr. O'Conor decided that the question of taking over the Home Department statistics should await further consideration.

52. By Home Department Resolution, dated 31st July 1896,[14] the statistical work of that Department was transferred to the Director General of Statistics on condition that he would deal with it in consultation with the Department and under its general control. In fact it was understood that the Director General had no authority to make any change in the form or substance of the returns without the previous approval of the Department. The work thus turned out to be of a mechanical kind consisting of the transcription of the figures from the Provincial returns in the form prescribed from the volume of Judicial and Administrative statistics (which was being issued by the Department since 1879). The Population Census was retained in the Home Department.

53. The next step forward was not taken till nine years later but one or two events of interest that took place in the interim may now be mentioned.[15] Firstly, with effect from 1st April 1901, on the initiative of Mr. O'Conor, the "Statistical Department" of the

---

[14] Commerce and Industries Department, C. & T. File No. 62/1907, March (e)40 page 5 last para.

[15] No. 2060 dated 19-4-1901, Finance & Commerce Department. Progs. No. 149-salaries etc.

Government of Bengal (consisting of 13 clerks and 7 compilers) which was dealing with the statistics of inland trade of the Province was transferred to the Director General's Office and a Bengal section was created in the Office. It was later on discovered, however, that this step contravened an earlier ruling (January 1890) by the Government of India that the registration and review of inland trade statistics was the business of the Provincial Agricultural Department; and, in October 1907, when the partitioned Province of Eastern Bengal and Assam made a similar request to the Government of India, the latter turned it down on the ground that there was no valid reason for a further infringement of the important principles enunciated in 1890.[16] Mr. Noel-Paton (Director General of Commercial Intelligence) on whose advice the Government of India refused the request, was of the view that the proper function of his Office in regard to statistics was essentially that of a central establishment controlling the system and combining the Provincial returns while holding the Local Governments responsible for their accuracy. The Bengal Section of the Director General's Office was, however, not retransferred to the Government of Bengal till 1923.

54. When Mr. O'Conor retired (1902) he was succeeded as Director General of Statistics by Mr. J. A. Robertson of the Financial Department.

55. It would appear that towards the end of 1903 an attempt was made in the Financial Department to estimate the total annual expenditure on the collection of statistics and that, side by side, there was a proposal to levy a "statistical" tax to cover this expenditure. But this is all that is known about the whole thing; presumably the scheme fell through.

Appointment of a Director General of Commercial Intelligence

56. We must now come down to March 1905 when Lord Curzon's Government with the previous sanction of the Secretary of State, constituted a new Department of Commerce and Industry and assigned to this Department the control of 'statistics' which was previously exercised by the Department of Finance and Commerce.[17] The new Department took over from the Revenue and Agricultural Department, the Offices of the Reporter on Economic Products, the Director, Geological Survey of India, the Chief Inspector of Mines and various types of data collected by the latter agencies. The responsibility for the co-ordination of inland trade statistics was also transferred to the new Department. The Home Department gave up "factories" but retained the Population Census.

57. It was decided at the same time that a Director General of Commercial Intelligence should take the place of the then existing Director General of Statistics; the new Officer was to act as liaison between the new Member for Commerce and Industry (Mr. Hewett) and the mercantile public. Mr. Noel-Paton (who was at that time Secretary of the Bombay Chamber of Commerce) was appointed to the post and in order to enable the new Director General to discharge his functions efficiently, a competent officer was to be appointed Assistant Director of Statistics so as to relieve the Director General of all but the more important work.

58. At about this time proposals for the appointment of a Committee to revise the trade returns had reached an advanced stage and Mr. Robertson (the outgoing Director General of Statistics) was to sit on it; and it was considered inadvisable to make over the

---

[16] File No. 21 of 1907, Department of Commerce & Industry.

[17] Department of Commerce & Industry, General B Progs., No. 25—30 April, 1905.

statistical work to Mr. Paton until after the revision of the returns.[18] Mr. Robertson was therefore placed on special duty for six months for the statistical work after the abolition of the post of Director General of Statistics with effect from the 27th of February 1905.

59. The Committee of Secretaries who had previously examined the question of the appointment of an Assistant Director of Statistics had recommended the selection of a ministerial Officer for the post, but in making this suggestion they had not taken into account the possibility of the post of Director General going to a man with commercial qualifications, but with no administrative experience.[19] Mr. Noel-Paton, the new Director General had no knowledge of administrative methods. It was therefore felt necessary that in the interests of Departments which were in control of the administrative statistics compiled in the Office of the Director General, the post of Assistant Director should be held by a man of wider experience than at first contemplated. In view of the larger responsibilities that would thus devolve upon the Assistant Director, the Government of India were of the opinion that the scale originally recommended, namely Rs. 700–50–1,000 per month was too low and they accordingly sought the sanction of the Secretary of State towards the end of March, 1905 for fixing the pay of the post at Rs. 1,300 so as to enable them to obtain for a period of three years the services of a competent officer who would be a junior member of the Indian Civil Service, possessing sufficient knowledge of the details of district administration and would be able to deal adequately with questions arising in connection with the compilation of administrative statistics.

60. The Secretary of State, in his despatch, dated 19th May 1905 agreed that the new Director General was inexperienced in administrative and official methods and that he should be given special assistance in this direction for a time, and sanctioned the proposal. At the end of the three years, however, the Director General should become conversant with administrative practice and should then be able to dispense with special assistance. He therefore directed that after that period the appointment should go to a non-I.C.S. Officer on the scale originally proposed.

61. Mr. Noel-Paton was not satisfied. He said that expert knowledge in statistical work could be acquired only by long experience and when it had been acquired it could not be "bequeathed" to a successor.[20] He protested that "the prospect of succession of short service novices, fitting through an office in which ripe experience was indispensable, postponed all hope of real efficiency." The Government of India had no choice, however. Mr. D. Shearme of the Indian Civil Service who was Private Secretary to the Governor of Burma was accordingly appointed to the post and he assumed charge on the 7th of August, 1905. He had the advantage of three weeks training with Mr. Robertson before the latter reverted to Finance Department.

62. This arrangement did not, however, work satisfactorily and hardly nine months had elapsed before Mr. Noel-Paton again raised the question and urged on the Government of India the expediency of appointing an Officer who might reasonably be expected to stay for a long time and become a real expert in Indian statistics. He said that the tenure of the post by a member of the Civil Service was subject to the disadvantage that such an officer could not be expected to make his career in statistics and that he would inevitably

---

[18] Letter No. 1625, dated 17-3-1905, Finance Department.

[19] Department of Commerce & Industry Progs. No. 34–35, May 1906, Commerce & Trade.

[20] File August No. 40/1905 C & T. (A), Commerce & Industry Department.

look elsewhere. Mr. Paton added that if one could find in India the man with the exceptional qualifications needed and who was not too old to learn the work, he would be preferable, but failing him it was better to obtain an expert, preferably a Chartered Accountant, from England rather than accept bad material for, in his opinion, district experience helped little in statistics.

63. Mr. Paton had made out a pretty strong case but the Government of India felt that if a really expert statistician was required for the immediate control of the Department it was questionable whether there was any advantage in subordinating him to the Director General of Commercial Intelligence. In the circumstances the matter was allowed to lie over and when Mr. Shearme took leave towards the end of July 1906 for three months, Mr. C. W. E. Cotton, I.C.S. of the Madras Service was appointed temporarily to act for him.[21] When Mr. Shearme was recalled by the Government of Burma a year later, Mr. Cotton took his place.

64. In the meantime, the question of appointing an expert statistician had been reopened in December 1906 and had at the suggestion of the Finance Department, been referred to the Home, and Revenue and Agricultural Departments who were also concerned in the matter.[22] The two Departments had no objection to the appointment provided "the general direction and control" of their statistics continued to rest with them that is, the new Officer was not empowered to make any material alteration in the forms prescribed by the Department or "to call for any but the merest routine information from, or issue orders to Local Governments without consulting the appropriate Department and obtaining their previous approval." But both the Departments agreed that a great improvement could be effected in statistics by having a trained statistician to deal with them. Mr. Risley, (later Sir) the Home Secretary (previously Census Commissioner) also pointed out that the expert would be of great assistance in solving the statistical problems arising out of the Census. Finance Department were doubtful whether the requisite qualifications for the post would be found in a Chartered Accountant; moreover, in their view, the supervision of the collection of statistical data required some administrative experience as well but the Finance Member was not prepared to oppose the proposal. The new Commerce Member (Mr. Finlay), however, after a discussion with Mr. Noel-Paton was of the opinion that there was no real urgency in the matter and he decided that action should be deferred to January, 1909.

65. The matter was ultimately raised with the Secretary of State in a despatch, dated 16th January, 1908 issued from Finance Department.[23] In the despatch the Government of India stated that after careful consideration they had arrived at the conclusion that if the Director General of Commercial Intelligence was to carry out efficiently the most important of the functions for which his Department was created, it was essential to associate with him a competent and fully trained statistician. The latter would, after a short experience of office routine and procedure, be able to assume effective control of the statistical work of the Office and would relieve the Director General of this branch of the work. The permanent nature of the new appointment would afford the holder every inducement to

---

[21] August 6–19, File 129/1907 C. & T. (A) Dept. of Commerce & Industry.

[22] March No. 40 File 62/1907 C. & T. Department of Commerce & Industry.

[23] File No. 15/1908 C. & T.(A) June 1–5, Department of Commerce & Industry.

master thoroughly the subject of Indian Statistics. The Government of India, in conclusion, requested the sanction of the Secretary of State for the following proposals :

(a) that an expert statistician with the designation of Director of Statistics be appointed to the position then held by the Assistant Director of Statistics;

(b) that the maximum pay of the appointment be Rs. 1,500 per month, the initial pay being decided by reference to the qualification of the person appointed;

(c) that the holder of the appointment should ordinarily control the statistical work of the Commercial Intelligence Department but that the actual appointment of the work would rest with the Director General of Commercial Intelligence; and

(d) that he is selected by the Secretary of State in England.

66. The Secretary of State (John Morley) rejected the entire proposal. In his reply dated 13th March 1908,[24] he pointed out that the proposal to appoint a Director of Statistics with a maximum salary of Rs. 1,500 a month was, in effect, a proposal to review the appointment of the Director General of Statistics side by side with that of the Director General of Commercial Intelligence and he said that he was not convinced of the desirability of relieving the Director General of Commercial Intelligence of the statistical work of the Department. In his opinion, a trained Statistician of the type contemplated by the Government of India was simply not required and he hoped that it would be possible to select (on the scale of Rs. 700–50–1,000 a month) from the staff of Accounts Department, or of the former Statistical Department, an Officer (who should be an Indian, if possible) fully qualified, to deal satisfactorily with the statistical work of the Department under the supervision of the Director General of Commercial Intelligence.

67. In the absence of Mr. Noel-Paton who had gone on furlough, the selection of a suitable non-I.C.S. officer was held up for some time and the post of a Civilian Assistant Director of Statistics (then held by Mr. G. T. H. Bracken, I.C.S. of the Madras Service) was extended for a period of six months from 7th August, 1908.

68. When the extension of six months specially sanctioned by the Secretary of State in August, 1908 for the post of Civilian Assistant Director of Statistics on Rs. 1,300 a month was about to expire, the Government of India in accordance with the previous instructions of the Secretary of State made an effort to find a suitable non-I.C.S. officer preferably an Indian, on Rs. 700–50–1,000 a month[25] and Mr. E. Sequeira of the Bombay Customs Service was appointed to the post in February, 1909; and he was confirmed as Assistant Director of Statistics in January 1910.

69. There were several other developments of statistical interest during the same period. In January, 1905 a small Committee consisting of Messrs. Morshead and Arbuthnot of the Customs Department and Mr. Robertson, Director General of Statistics was appointed to revise the statistical publications of Government relating to trade and cognate subjects, with the object of rendering them more accurate and more generally useful.[26] After visiting the principal Indian ports and consulting the Customs Authorities and the Commercial

---

[24] File No. 15/1908 June C. & T.(A).

[25] Department of Commerce & Industry, File No. 10 (Commerce & Trade) of 1914-15, February (A) 1–18.

[26] File No. 3/1908, August, Stat. (C), June, Stat. (A).

community in each, the Committee submitted its report in August, 1905 suggesting various improvements in the method of registration and compilation of the foreign, coasting, rail and river-borne and land frontier trade of India; and the Government of India in their Resolution dated 2nd March, 1906 accepted nearly all the recommendations and such of these as required action at ports were forwarded to the Local Governments who were at that time the primary authorities for the collection of these statistics. An interesting appointment made in the Office of the Director General of Commercial Intelligence was that of an Auditor (Travelling Statistical Inspector) on Rs. 240 per month. It was the duty of this official to visit the different Railway Audit Offices in turn and in each office check a set of weekly or fortnightly returns taken at random.[27] Previously, the figures supplied by the Railway companies were subjected to little check and there was no guarantee of uniformity in the systems adopted in the different Railways.

70. In July 1905, Mr. Holland (later Sir Thomas), Director of the Geological Survey of India proposed to the Commerce and Industry Department that, the responsibility for the publication of the annual statistics of mineral production should be transferred to him, of course without disturbance to the then existing system of collecting figures from the Local Governments.[28] The returns used to be compiled originally in the Statistics Branch of the Revenue and Agricultural Department. In 1892 the work was given to the Reporter on Economic Products. In 1899 the publication was split into two, the one being annual and the other quinquennial; and the former was transferred to the Director General of Statistics. In 1903 the quinquennial review was handed over to the Director of the Geological Survey of India and the latter in course of time not unnaturally desired that the annual publication should also be given to him. Mr. Paton and Mr. Robertson had no objection to this but according to the new arrangement, the Director General of Commercial Intelligence was to receive the figures from the Local Governments and forward them to the Geological Director for consolidation and publication.

71. In November, 1905, an establishment consisting of 15 persons (including a Personal Assistant to the Director General of Commercial Intelligence on Rs. 500 a month) was sanctioned temporarily for a period of two years for the purpose of bringing out the Indian Trade Journal.[29] The first issue of the Journal came out on 5th April, 1906. In June, 1907 the decision was taken to place the staff on a permanent footing but the Secretary of State, refusing to sanction extra staff for the Journal clearly laid down that the Journal should confine itself to fully authenticated trade intelligence and statistics and that all expressions of opinion should be cut out.[30]

72. In January, 1907, Mr. Noel-Paton pointed out to Government the several difficulties encountered in producing miscellaneous volumes like the Financial and Commercial Statistics and Judicial and Administrative Statistics, e.g., undue delay in publication, uneven pressure on his Office and rush in the press.[31] He therefore suggested that the matter contained in the two volumes might with advantage be split up into nine parts of 95 pages each, the statistics being so re-arranged that each part would contain only kindred subjects. The

---

[27] File No. 3/1906-September (A) 6–7, Commerce and Industry Department, Stat.

[28] File No. 128/1905, Geo.& Minerals, July 1.

[29] File No. 64/1907, C. & T., July A 6–8, Commerce & Industry Department.

[30] File No. 64/1907 C. & T. March A 19–20 of 1908, Commerce & Industry Department.

[31] File No. 10/1908, April–Stat. (C).

three Departments concerned, *viz.*, Finance, Commerce and Industry, and Home readily agreed to this. They also concurred with Mr. Paton's proposal made a year later that, in carrying out such other alteration as might appear necessary in arrangement and presentation from time to time, he should be allowed a free hand. However, the complete exclusion of any old table or the introduction of a new one and alteration in tables issued as appendices to Government Resolutions had to be departmentally approved.

73. In July, 1907 arising out of a complaint from the Government of Bombay against the tone of a letter written to them by the Director General of Commercial Intelligence, the Department of Commerce and Industry made it clear that it was never the intention of Government that Mr. Paton should be constituted the sole authority (although he might be regarded as the central authority) for receiving and answering enquiries and that it was undesirable that delay should be caused by insisting that information not readily available should first be collected in central office before being communicated to the enquirer.[32]

74. Early in 1910 the Government of India decided that a detailed investigation should be undertaken of the problem of the rise in prices which had by then assumed considerable importance and in March of that year, Mr. K. L. Datta, a senior and experienced officer of the Finance Department was appointed to conduct the enquiry assisted by Mr. Findlay Shirras of the Indian Educational Service and Mr. Gupta of the Finance Department. The Indian States and Burma were to be excluded from the enquiry owing to paucity of statistical information.

75. Mr. Datta and Messrs Shirras and Gupta toured extensively in India with a view to collecting reliable statistics of prices and wages on the spot and examining the machinery employed in the primary collection of the statistics published by the Director General of Commercial Intelligence. They interviewed both officials and non-officials and collected a mass of statistical material on wholesale and retail prices, wages of skilled and unskilled labour (both in rural and urban areas), acreage and outturn of crops, rainfall, population, external and internal trade, coinage and currency and banking, freights, transport, etc.

76. Mr. Datta's report was received by Government in 1913 and the complete subsidiary volumes of statistics were ready by April, 1914. The Government of India Resolution on the subject issued in October, 1914 stated that the Report with its statistical appendices constituted a complete survey of the progress achieved in the previous 22 years and that it must be recognised as a very valuable contribution to the economic and financial history of India.[33]

Appointment of a director of statistics

77. Early in 1912, after the transfer of the Headquarters of the Government of India from Calcutta to Delhi, the question of reorganization of the Department of Commercial Intelligence came up for immediate consideration. The experience of the working of the Department had clearly demonstrated the inadvisability of combining under the charge of a single officer, duties coming under two distinct categories, viz., (1) collection and publication of statistics, and (2) the distribution of commercial intelligence. Accordingly after a careful examination of the whole question, the Government of India addressed the

---

[32] Department of Commerce & Industry, File No. 73 of 1907—Sept. C. & T.(C).

[33] Finance Department, Resolution No. 1614/F, dated 24th October 1914.

Secretary of State in July, 1913 requesting his sanction for the separation of the statistical work from the other work of the Commercial Intelligence Department and for the appointment of a Director of Statistics on Rs. 1,200-50-1, 750 a month, to be located in due course at the headquarters of Government.[34]

78. The Secretary of State agreed that the work of compilation and publication of statistics merited the individual attention of an experienced officer and gladly sanctioned the porposal. At the same time the Secretary of State asked the Government of India to consider whether the duties of the Director General of Commercial Intelligence could not with advantage be transferred to suitable Provincial Agencies and he said that if this was found possible, there might not be any further need for Director General. He could, therefore, sanction the extension of the latter post for a period of three years only.

79. The post of Director of Statistics was created from 1st April, 1914 and Mr. Findlay Shirras of the Bengal Educational Service was appointed to the post, with His Excellency the Viceroy's sanction, for a period of five years in the first instance.[35] The post of Assistant Director of Statistics was abolished with effect from the same date but Mr. Sequeira was also to help the Director of Statistics "as far as possible."

80. Mr. Noel-Paton had been informed in the middle of February of the proposed appointment of the Director of Statistics and he had been asked to suggest the division of functions between him and the new Director of Statistics keeping in mind the work that was being done by the Director General of Statistics before 1905. Within three days Mr. Noel-Paton sent his suggestions to the Government of India but he went on leave immediately thereafter and Mr. A. H. Ley, I.C.S., was appointed to officiate for him. Mr. Paton's suggestions as amended by Mr. Ley and finally accepted by Government were as follows :

(a) The Review of Trade should be drawn up and published by the Director of Statistics, who would consult the Director General of Commercial Intelligence on such portions of the report as were not purely statistical before submitting it to the Department of Commerce and Industry.

(b) Revision of tariff valuations should be entrusted to the Director General of Commercial Intelligence.

(c) Crop forecasts should be left to the Director of Statistics.

(d) The Indian Trade Journal (whose future was still undecided) would be under the control of the Director General of Commercial Intelligence who would be entitled to call on the Director of Statistics for such statistical matter as he desired to be embodied therein. The confidential Catalogue of Indian Manufactures remained with the Director General of Commercial Intelligence.

81. Apart from the above, all well-defined statistical work that was being done by the Department was to be taken over by the new Director. The Rate Lists which were being compiled with a view to enabling purchasing offices in India to compare prices of articles in India and England were also to be transferred to the Director of Statistics.

82. As for the division of establishment, it was decided that in view of the uncertainty that prevailed in regard to the future of the Commercial Intelligence Department, it was best for the Director General of Commercial Intelligence and the Director of Statistics

---

[34] Govt. of India, Despatch No. 186, Finance of the 17th July, 1913.

[35] Department of Commerce & Industry, File No. 10 of 1914-15, February A 1-18.

to manage together with the then existing staff for at least six months although the close nexus that had existed between the Commercial Intelligence and Statistics sections was to be gradually severed. The question of handing over the Trade of Bengal section to the Government of Bengal, was also raised but the Local Government did not exhibit any anxiety to take it back and further, there was no particular urgency in the matter as the transfer of the Department of Statistics from Calcutta to Delhi was not expected to come off soon.

83. Mr. Shirras's Office consisted of about 100 hands divided into ten sections, *viz*., (I) General (Registry, issue and records) (II) Prices and Freights (III) Wages (IV) Judicial and Administrative Statistics (V) Rail and River-borne Trade of India (VI) Trade of Bengal (VII) Sea-borne Trade (VIII) Frontier Trade (IX) Agricultural Statistics and Forecasts (X) Commercial, Industrial and Financial Statistics.

84. The last section dealt with the statistics of joint stock companies, industrial establishments and banks.

85. Below the Director there were only two Superintendents, each Superintendent being in charge of a number of Sections. Apart from the paucity of staff there were other faults in the working of the statistical machine such as cumbrousness of the method of disposal, excessive noting, distance from the Headquarters of Government, and round about correspondence.

86. In April 1918, Mr. C. E. Low (Secretary, Commerce Department) inspected the Office while in Calcutta with the Indian Industrial Commission and in his report he said that the Office was well supervised and that the outturn of clerks was fully as much as could be expected from a Calcutta Office but that the issue of returns was in some cases very much belated.[36].

87. Mr. Shirras's term of office was nevertheless marked by a number of improvements in the quality and representation of current official statistics. It was in his time that the first concerted step was taken to improve the statistics relating to Indian States by calling a Conference of Ruling Princes and Chiefs.[37] Mr. Shirras also brought out three important special publications, namely :

(i) A Manual on the Preparation of Crop Forecasts in India.

(ii) A Collection of Orders Relating to the Registration and publication of Statistics Relating to the Sea-borne Trade ; and

(iii) a Guide to the Statistical publications of the Government of India.

88. In a memorandum prepared for the Procedure Committee under Sir Hubert Lewellyn-Smith's chairmanship in October 1919. Mr. Shirras made the following suggestions for remedying the defects noticed in the Department:[38]

(1) The gazetted staff should be strengthened by the immediate appointment of an Assistant Director and gazetting the two Superintendents.

(2) The non-gazetted staff should be strengthened in the higher division and recruitment in the lower divisions discontinued.

---

[36] Department of Commerce & Industry Statistics 1920, January (B), 1.2.

[37] Department of Commerce & Industry 1918/Stat./September, 8.

[38] Progs. No. 13 (Statistics) of February of 1920, Dept. of Commerce & Industry.

(3) Greater use should be made of labour saving devices.

(4) The Director of Statistics should be made Deputy Secretary (Statistics) in Commerce Department and should be located at the Headquarters and should be permitted to dispose of all cases except those relating to expenditure beyond a certain amount.

(5) Noting should be curtailed.

(6) Correspondence on Statistics between India Office and the Director of Statistics should be direct and not through the Department of the Government of India concerned as this meant delay.

89. Mr. Shirras also desired that the Census should be transferred to the Department of Statistics in so far this was practicable.

Report of the Indian Industrial Commission

90. The Industrial Commission under the Chairmanship of Sir Thomas Holland had, however, reported against any extension of the Statistical Department.[39] The Commission after taking a great deal of evidence on the suitability of the then existing system of collecting and distributing statistics of commercial intelligence under the Director of Commercial Intelligence respectively, arrived at the conclusion that where a specialised Department existed, it should handle the statistics of interest to that Department, as in the Railway and Mineral statistics. Unless this was done, they said, irrelevant data would be collected and essential facts would be overlooked.

91. The Commission was of the view that the Director of Statistics should be a compiling officer only. In other words, his relation with statistics should be merely arithmetical, that is, he should not be allowed to comment on them. As an example, the Commission cited the case of crop forecasts, the preparation of which was a highly technical business and expressed their disapproval of the existing practice whereby the Director was responsile for amending the forecasts of Provincial Officers. Such a duty, would, in the opinion of the Commission, be more appropriately performed by a responsible officer having general experience of Indian agriculture.

92. The Commission also recommended the constitution of a separate Department of Industries and a Director of Commercial and Industrial Intelligence under the general control of that Department but they were clear that the Director of Statistics should remain under the control of the Commerce Department of the Government of India.

(The Commission agreed with the suggestion of the Government of India that the Indian Trade Journal should be discontinued and in its place, occasional bulletins on subjects of interest should be issued from time to time).

In the circumstances the term of office of the Director General of Commercial Intelligence, then Mr. Lindsay, was being extended from year to year.[40].

93. The Government of India rejected the recommendation that the Commercial Intelligence Department should be transferred to the proposed Department of Industries. At the same time after consulting Provincial Governments and Chambers of Commerce they

---

[39] Report of Indian Industrial Commission 1916–18, pages 120--25.

[40] 1919 Jan, A 1–3 Commerce Intelligence Department of C. & I.

submitted to the Secretary of State (Mr. Montagu) a scheme for a considerable expansion of the staff of the Commercial Intelligence Department including an increase in the pay of the Director General of Commercial Intelligence.[41] They also stated in the despatch that they had for the time being dropped the idea of recruiting commercial men for the service and were on the look out for a suitable I.C.S. Officer. In regard to the Indian Trade Journal, the Government of India were now not sure that it should be discontinued although they themselves had suggested it in 1916. In his despatch dated 20th January, 1921, the Secretary of State agreed to these proposals.

94. The Government of India accepted the suggestion that the Director of Statistics should remain as a separate entity under the Department of Commerce. But the views of the then acting Member for Commerce and Industry (Sir Thomas Holland) are worth quoting in full :[42]

"The Director of Statistics now deals with returns for industries, agriculture, public health, education, justice and dozens of things that do not come under the Commerce Department. The raw materials from which he constructs his tables should, when possible, pass through the expert sieve of the appropriate specialist department, as in the matter of minerals since 1904; and the specialist, not the statistician, should point the lessons; apart from the cause of accuracy, the handling of statistics suggests lines of action for the specialist departments, which may or may not, but should generally, issue separately their reviews of the figures. Statistics are never accurate, but when passed through the specialist sieve, the residual errors are relatively unimportant; without this process of refining they are as dangerous as Herbert Davis clock.

"There is room for a special department in which the chief statistics are gathered together suitably for public reference; and although the chief fraction by bulk of the statistics will always relate to trade, the Director of Statistics need not be under the administrative control of the Director General of Commercial Intelligence. The latter, however, is the appropriate specialist for reviewing commercial statistics; the present practice by which the Director of Statistics reviews the trade returns is a mistake."

95. The Imperial Statistical Conference which met in London early in 1920 adopted an important resolution on the organization of Statistical work. The resolution said that "the functions of no statistical office can be exercised properly when limited to the initial work of collecting and compiling figures. The Conference is agreed that the prescription of so narrow a function has resulted in the past in a tendency on the one side to an unintelligent tracing of undigested statistical material and on the other hand to a failure to render information which is of value readily available for public use".[43] The Report of the Conference emphasised the fact that the arithmetical work of a Statistical Office constituted only the preparatory and less important stage of the work of such an Office and that if this principle be ignored the more important services which through statistical work is capable of rendering to the community will be sacrificed.

96. The Conference thus contemplated that the Statistical Officer should not only compile the figures but also publish them with some form of critical analysis. This was

---

[41] Finance Department Despatch No. 431, dated 10th Sept., 1920.

[42] Department of Commerce (Commercial Intelligence), progs. July 1923, Nos. 1–26, page 15.

[43] Commerce Department Statistics, progs. 1–2, part B, December 1922 and K.W.

clearly in conflict with the recommendation of the Indian Industrial Commission which had previously been accepted by the Government of India, *viz.*, that the Director of Statistics should regard himself as merely a compiling Officer. Mr. Shirras asked for a reconsideration of this decision but the Hon'ble Member for Commerce was unwilling to reopen the question.

97. In another resolution, the conference affirmed its opinion that the accuracy of statistical information secured by means of voluntary or optional returns was open to serious doubt. The conference, therefore, recommended that whenever a Government called for statistics on any subject or series of subjects it should confer the necessary statutory authority upon its officers. But when the question came up for consideration by Government of India, it was felt that this resolution went too far in the case of India; Government were of the view that there would be no harm in compelling well-organized industries such as jute mills, collieries and the like to supply specified information. This was already admitted in principle in the case of cotton but it would be unwise to undertake general legislation even of the comparatively limited scope of the English Bill and it was decided that for the time being it would be better to rely mainly on voluntary information.

98. It must be stated here that the taking away of the Review of Trade did not mean any diminution in importance of the post of Director of Statistics, for towards the end of 1921 the Government proposed an enhanced rate of pay on the scale Rs. 1,250-75-2,000 for the post. Before the sanction was received, however, the permanent incumbent of the post, Mr. Shirras had been appointed Director of the newly opened Labour Office under the Government of Bombay (from 1st April 1921) and Rai Bahadur D. N. Ghosh had succeeded him as officiating Director of Statistics; and the sanction was treated as cancelled. The Director General of Commercial Intelligence (Mr. Lindsay) was placed in general control of the Statistical Department. He was asked to write the Annual Statistical Reports in addition to the Review of Trade.[44]

99. By this time the financial situation had greatly deteriorated and the Governor General had, with a view to effecting retrenchment in every possible direction appointed a Committee under the Chairmanship of Lord Inchcape.[45] Departments were told that in preparing their cases "they should bear in mind the fact that certain activities, although eminently desirable in themselves, may be beyond the means of the Government of India in present financial conditions.

100. The axe fell heavily on the Department of Commercial Intelligence and Statistics. The post of Director General of Commercial Intelligence on Rs. 2,500-100-3,000 per month was abolished, but Mr. C. G. Freke, I.C.S. who was officiating in that post at the time was permitted to retain the title in the interests of continuity in the matter of correspondence with Commercial Departments in other countries. The post was, however, to be revived soon afterwards.

101. The Department of Statistics was amalgamated with the Commercial Intelligence Department and the Director of Statistics, while retaining his title, was placed on the footing of a Deputy Director of Commercial Intelligence on a pay of Rs. 1,000-100-1,500 per month. The Bengal Section of the Office was to be re-transferred to the Provincial Government with effect from 1st April 1923. The Director General of Commercial Intelligence was

---

[44] Department of Commerce & Industry, Progs. B No. 4-19, May 1921.

[45] Finance Department, O. M. No. 1333-F, dated 5th May 1922.

asked to examine the publications of the Statistics Department in consultation with Director of Statistics with a view to their simplification and curtailment. It was later resolved to discontinue the "Prices and Wages" and "Rail and River-borne Trade Statistics." The five volumes of the Statistics of British India were substituted by the single volume of the Statistical Abstract of British India and the corresponding India Office publication (which they were unwilling to give up in 1915) was discontinued.

ECONOMIC ENQUIRY COMMITTEE

102. Early in 1925, pursuant on Resolutions passed by both Chambers of the Central Legislature, the Economic Enquiry Committee under the Chairmanship of Sir M. Visveswarayya was set up to enquire into "the question of the adequacy of the statistical data available and the desirability and possibility of supplementing it and of undertaking an economic enquiry."

103. The Committee in their majority report which was submitted in August 1935 made a comprehensive survey of the available statistical material in regard to production, prices, etc. and came to the conclusion that if statistics in India were to be maintained on a satisfactory basis, all work relating to it should be co-ordinated and centralized as in the Dominions; in other words, the statistics of all Departments, both of the Central and Provincial Governments should come under the supervision of one central authority who would act as the adviser to Government in all statistical matters.

104. The Central Organization proposed was to consist of a Director of Statistics, whose appointment would be subject to the vote of the Central Legislature and two Assistant Directors, a third being employed for temporary periods of about four years each time in connection with the decennial Population Census so as to assist the Director of Statistics who would also take over the responsibilities of the Census Commissioner. The Committee also proposed the setting up of a Statistical Bureau in each Province under a Provincial Statistician possessing a high degree of local knowledge. Citing the old saying that there can be no statistics without legislation the Committee pointed out the need for placing the entire Organization on a statutory basis by enacting a Census and Statistics Act.

105. Among the other recommendations made by the Committee was one relating to the compilation of an Official Yearbook on the lines of the Dominion Yearbooks, with the co-operation of the Director of Public Information.

106. The Government of India appreciated the underlying object of the Report which was to improve the standard and extend the scope of the available statistical material but in view of the important financial and administrative considerations involved in implementing the recommendations, they decided in the first instance to invite the views of Provincial Governments on the proposals.[46]

107. At about the same time, the President of the International Statistical Institute sent a copy of the Resolutions adopted at the 16th session of the Institute held in Rome in 1925 and recommended their adoption by the Government of India.[47] In October 1926 the India Office asked for the views of the Government of India on these recommendations with particular reference to a census of industrial production.

---

[46] Finance Department, Resolution No. D/4299/F, dated 23rd Oct., 1925.

[47] File No. 527 A-C(2), Com., February 1927.

108. It was out of the question, however, that at a time when the statistical budget had been so badly cut down that even inland trade statistics had to be discontinued, any new expenditure on the collection of industrial statistics would be countenanced by the Government of India. The Commerce Secretary's reply to His Majesty's Under Secretary of State, however, merely stated that the Government of India proposed to defer consideration of the Resolutions until they had received and considered the views of Provincial Governments on similar recommendations made by the Indian Economic Enquiry Committee.

109. Provincial opinion was rather slow in expressing itself but the Government of Burma, who replied after a year saying that they would oppose any proposal to be handed over to the Central Bureau of Statistics, the control of the collection and compilation of Provincial statistics. While they had no objection to advice from the Central Bureau, they were firmly of the view that the Provincial Administrative Departments should be given freedom to accept or reject this advice.[48]

110. The reply from the Government of Burma proceeded to say that the Local Government considered it desirable to leave the existing agencies untouched but to subject the statistics collected by them to the expert scrutiny of a Provincial Director of Statistics who would be otherwise unconnected with these agencies. They therefore proposed to appoint an I.C.S. officer to this post for a period of three years if the Government of India had no objection. Replying in March 1927 the Government of India said that they had not yet completed examination of the question of creation of the Central Statistical Bureau and hence they could not commit themselves one way or the other at that stage.

111. Ultimately in their final orders on the Report which were issued on 28th August 1929 the Government of India stated the subject of Provincial Statistical Bureaus had definitely been superseded by the recommendations of the Royal Commission on Agriculture which were being examined in consultation with Local Governments. It was also stated that the question of the creation of a separate Department of Statistics was under Government's consideration.

112. In regard to the proposal for the compilation of a Yearbook it would appear from the Departmental notings that it was considered undesirable to enter into competition with private enterprise, namely, the Times of India Yearbook.

113. As early as 1924, the Government of Bombay had proposed to introduce a bill in the local legislature to provide for the statutory collection of statistical information for public purposes. Statistics was then a Central subject (Entry 39 in Part I of Schedule I to the Devolution Rules) but the Government of India readily gave the necessary sanction to the introduction of the Bill. They were at the same time doubtful whether the country as a whole was ripe for legislation of this kind.

114. A Bill was accordingly prepared in the Bombay Labour Office so as to provide for the collection of statistics relating in particular to prices of commodities, cost of living, rents, rates of wages, and hours of work, employment and unemployment and industrial disputes and generally, any other matter which the Governor-in-Council might by notification in the Bombay Government Gazette prescribe. On introduction in the local Legislative Council however, it met with a mixed reception and after a lengthy debate on the motion for first reading the Council decided that the Bill should first be circulated to the various Associations concerned and then be re-drafted in the light of the opinions received.

---

[48] File No. 1393-C, Commerce, April 1927.

115. Early in 1926 the Government of Bombay requested the sanction of the Government of India for their re-drafted Bill. They also wanted to know whether in view of the Report of the Indian Economic Enquiry Committee, the Government of India had any objection to the Provincial Government embarking on such legislation. But before the Government of India could make up their mind in the matter, the Government of Bombay had independently decided not to proceed with the Bill.

ROYAL COMMISSION ON AGRICULTURE IN INDIA

116. The Royal Commission on Agriculture which reported in 1928 also examined the general question of statistical organization apart from the other questions concerning agricultural statistics. They recommended the appointment of a well-qualified Statistical Officer at each Provincial Headquarters, who would have the status of an officer of the Superior Provincial Services and would be entrusted with the task of studying all aspects of the social and economic progress of the Province.[49] He would act as the Director of a Bureau of Statistical Information with a small staff of subordinate officials and would eventually be made responsible for co-ordinating and publishing all Provincial statistics. The wider the field a Statistical Officer was called upon to survey, the more opportunities he had for making those discoveries from the correlation of data derived from widely different sources which were the aim, and frequently the achievement, of all modern statistical work.

117. The Commission also pointed out that although the general tendency towards Provincial self-sufficiency must be recognized in the statistical as in other fields, it was nonetheless true that there were important statistical services which could be provided only by a Central agency. Facility of comparison between provincial statistics was essential to their full utility and the uniformity necessary for comparison would be most easily secured if machinery for coordination existed. The Commission were therefore in favour of reconstituting the Department of Statistics as a separate department. The Commission were, however, unable to agree with the Economic Enquiry Committee that a large central organization should be created. All that would be required at first would be the engagement of a first-class statistician of recognized competence who, with his staff would be appointed and paid by the Government of India. The Officer might be styled the Director of the Central Bureau of Statistical information. The Officer selected must be of sufficient standing to make his advice acceptable not only to the Provincial statistical officers but also to the business world and the informed public. Provision should also be made in the Central Statistical Organization for Boards which would advise on the publication of statistics and their periodical revision; and representative leading economists, scientists, and business men would find a place on these boards as well as officials of the Departments interested. For, the Commission stated, "the whole basis of statistics in India urgently requires broadening. It should rest not on the work of a few government officials, however able, but on the support of the informed public, and through them, on the recognition by the legislatures and by the general public that modern statistical methods are in a position to make an indispensable contribution to the successful development alike of scientific agriculture and of social administration."

118. The Commission expressed itself against general censuses of production or large schemes of economic investigation into the condition of the population undertaken with the mistaken notion that the mere accumulation of facts will of itself serve a useful purpose.

---

[49] Paras 538–39 of the Report.

ROYAL COMMISSION ON LABOUR IN INDIA

119. The Royal Commission on Labour in India which reported in 1931 had its own recommendations to make in regard to Statistical Organization. In addition to suggesting the establishment of Labour Offices at the various Provincial headquarters on the lines of the Bombay Labour Office and Boards of Economic Enquiry on the model of the Punjab Board, the Commission pointed out the misapprehension that prevailed in certain quarters to the effect that the collection of statistics was a luxury in which only rich countries should indulge. They said that it was on facts that policy must be built and so long as there was uncertainty as to the facts, there must be confusion and conflict regarding the aim. At the same time the Commission doubted if any extensive progress could be made without statutory powers. It was true that the Bombay enquiries had been carried through with the voluntary cooperation of millowners but it was unlikely that the same amount of success could be achieved everywhere by voluntary methods. Legislation for the collection of statistics regarding the economic condition of the people was in force in the majority of countries of industrial importance. The Commission therefore recommmended that legislation be adopted, preferably by the Central Legislature, enabling the competent authority to collect information from employers regarding remuneration, attendance and living conditions (including housing) of industrial labour, from merchants regarding prices, from landlords regarding rentals etc. In the view of the Commission there was no room for undue fear regarding the possible abuse of such powers by Government Departments and in this connexion they proposed the setting up of a statutory body in which representatives of employers, labour, and Governments would meet regularly in conference. A body of that kind when it was satisfied in regard to the need for compulsory powers and the discretion that would be shown in using them, would be able to give the necessary authority to officers. It would naturally be a condition that individual returns should not be published or disclosed without express permission.

120. A tangible outcome of the recommendations of the Royal Commission on Agriculture in regard to statistical improvement was the setting up of a statistical section in the newly opened Imperial Council of Agricultural Research. At about the same time, it was decided to revive the publications relating to statistics of internal trade. A scheme for the collection of statistics of industrial production on a voluntary basis was also put into effect.[50] The scope of the available wholesale price statistics was enlarged. Apart from these, there was nothing of importance during the period to report.

THE ESTABLISHMENT OF A STATISTICAL RESEARCH BRANCH

121. In response to public demand, His Excellency the Viceroy Lord Willingdon in an address to the Legislative Assembly early in 1933 forecast the initiation of a programme of planned economic development for India.[51] The success of such a programme was, however, dependent on the existence of a trained organization for the continuous analysis and interpretation of economic statistics, which India did not possess at that time. Government were thus faced with a situation which called for the immediate establishment of such an Organization. They accordingly decided to establish at their headquarters a Statistical Research Bureau under the control of the Director General of Commercial Intelligence and

---

[50] File No. 1762-C(10), Commerce A, August 1931.

[51] File No. 7(C)-(2) of 1933, Commerce A.

Statistics. At one stage, it was also proposed that the Statistical Branch of the Department of Commercial Intelligence and Statistics should be moved to Delhi but when the Director General (Dr. Meek) submitted the necessary detailed proposals, it was realised that difficulties would arise in regard to the provision of office and residential accommodation for the staff to be transferred. It was therefore decided that the move need not take place immediately. The question was no doubt re-opened some years later but it had to be dropped again, this time owing to difficulties connected with printing arrangements.

122. As the Director General would necessarily have to tour greatly, it was thought necessary to have at the head of the new Research Branch a Deputy Director of Statistical Research and the post was filled by an Officer already in Government Service. The Statistical Research Branch was later merged in the Office of the Economic Adviser.

Bowley-Robertson enquiry

123. The next important landmark in the history of statistical development in India was the Bowley-Robertson enquiry, in 1934, into the possibility of an Economic Census of India. Messers Bowley and Robertson covered a very wide ground and in addition to making detailed suggestions in regard to a Census of Production and the measurement of the National Income, they made a series of recommendations in regard to statistics of production, prices, wages and profits. They also recommended the compilation of a Guide to Current Official Statistics for the Centre and for the Provinces. This last recommendation has since been given effect to at the Centre by the Economic Adviser to the Government of India, but, in the Provinces not much headway was made, presumably owing to lack of a similar Organization.

124. The two experts further recommended the appointment of a permanent Economic Staff consisting of four, including a Director of Statistics at the Centre with a considerable degree of independence, and with duties identical with those specified by the Economic Enquiry Committee of 1925, namely, the organization of the Census of Population and the Census of Production, as well as the co-ordination of all Central and Provincial statistics. They also envisaged the appointment in each Province of a whole-time statisticians, independent of departmental control, who would cooperate with the Central Director of Statistics in every possible way.

125. As a result of the recommendations contained in the Bowley-Robertson Report, the Government of India decided in June 1936 to set up a Central Statistical Organization at headquarters at the earliest possible moment.[52] But this organization was to remain separate and distinct from the Economic Intelligence Organization which it was also proposed to create at New Delhi under an Economic Adviser to the Government of India and whose duties would be "the collection and study of information bearing or likely to bear on India's economic welfare and the supply of information and advice to the Government as a whole and to the Executive Departments as required." The Statistical Research Branch already in existence at Headquarters was to form part of the latter Organization and not the former.

126. It was also proposed once again that the entire staff then employed on statistical work in the Department of Commercial Intelligence and Statistics should be tranferred from Calcutta to New Delhi and should form nucleus of the Central Statistical Organization.

---

[52] File No. 7-C(10), 35 Comm. A Sept. 1938.

127. It turned out later, however, that owing to practical difficulties in the way of transferring from Calcutta to Delhi the work and staff connected with the printing of Statistical publications and the heavy financial commitments this would entail, the Government of India decided to abandon altogether the scheme for creation of a Central Statistical Organization. The Economic Intelligence Organization under an Economic Adviser Dr. (later Sir) T. E. Gregory had,on the other hand, been set firmly on its feet early in 1938 and had, as proposed, absorbed the Statistical Research Branch at headquarters.

DEVELOPMENTS DURING WORLD WAR II

128. The outbreak of the war in 1939 and its later transformation into a World War threw on the Government of India a two-fold responsibility for the execution of programmes of actual military operations including the procurement of supplies for the forces and the co-ordination and rationing of civil needs. The need for more and more statistics began to be keenly felt by the various Departments of Government for the effective discharge of their new obligations and a great mass of data was thus collected. For example, the Industries and Civil Supplies Department, which was specially created for the purpose of regulating supplies of various articles available for civilian consumption, compiled in connection with the administration of controls, continuous information relating to essential articles. As experience showed that in the absence of statutory powers, completeness and accuracy could not be obtained in regard to data on industrial output, legislation as planned by Dr. John Matthai, Officer on Special Duty in the Commerce Department in the early stages of the war period was enacted in the shape of the Industrial Statistics Act of 1942. Procurement and output statistics over a wide range of articles were collected by the Purchase Sections and Directorates of the Supply Department. The Department of Food garnered valuable data concerning prices, stocks and movements of food products as well as essential information relating to urban rationing. The War Transport Department which was also formed during the war collected new data relating to motor vehicles. The Railway Board expanded its statistics so as to show the extra strain on the railways. The Department of Commerce obtained detailed statistics pertaining to import and export trade controls.

129. Additional information in regard to acreage, production and prices of agricultural commodities and livestock was gathered by the Education, Health and Lands Department. Resulting from the deliberations of a small committee under the Chairmanship of Sir Pheroze Kharegat, a pilot scheme of crop cutting experiments relating to the principal crops was evolved in 1943 by the Statistical Adviser to the Imperial (now Indian) Council of Agricultural Research. The Department of Labour initiated a system of retail price index numbers for rural centres in various parts of India and made considerable progress in the matter of compilation, on a uniform basis, of working class cost of living index numbers for important industrial centres in India. The Office of the Economic Adviser to the Government of India was authorized to undertake a comprehensive enquiry into the family budgets of middle class employees of the Central Government.

INTER-DEPARTMENTAL COMMITTEE ON OFFICIAL STATISTICS

130. With all these, however, a number of gaps began to exhibit themselves and the problem of filling these gaps and improving the then existing organization for the collection of statistics led the Government of India early in January 1945 to set up a fully representative

Inter-Departmental Committee under the Chairmanship of Sir Theodore Gregory to review the position and suggest possible lines of reform. The terms of reference to this Committee were :

(i) to consider the statistical material available to Government and to make recommendations with particular reference to filling up the gaps;

(ii) to suggest such additional statistical material as may be required in the light of the probable needs of the Government and the country in the post-war period;

(iii) to review the existing organization by which the material is supplied and to suggest any improvement or changes considered necessary.

The Committee continued its deliberations over the whole of 1945 and before it reported, the war had ended and the focus of attention had necessarily shifted from war efforts to reconstruction. On re-examination, the Committee found that although a large portion of the mass of material collected during the war, e.g., statistics, of output and utilization of electricity, was of continuing utility, a good deal of new work still remained to be done. Accordingly, it made a number of detailed recommendations relating to the compilation of a comprehensive index of industrial production, employment and wage rates, price indices, trade classification, volume of retail sales, statistics of housing and building construction, and the output of cottage industry. The Committee laid special emphasis on the setting up of machinery for the continuous compilation of estimates of the National Income and Balance of Payments and of a durable organization for the Population Census. [The last of these items had also been touched upon by the Health and Development Committee and the Population Data Committee ; in fact recently a firm step forward in the direction was taken in the form of the appointment of a Registrar General of Population and Vital Statistics in the Ministry of Home Affairs and the setting up of a Population Advisory Committee as a standing institution.] Amongst the purely organizational recommendations of the Inter-Departmental Committee was a scheme for the decentralization of primary responsibility for departmental statistics coupled with the formation of a Central Statistical Office by transfer from the Department of Commerce to the Cabinet Secretariat of the co-ordinating functions and staff, the institution of a Statistical Cadre, the establishment of Statistical Bureaus at the headquarters of Provincial and State Governments and the preparation of an overall Statistics Act for the entire country.

With the establishment of the National Government in India, the functions of the Government expanded in many ways and the whole subject had to be reviewed once again.

Paper received : July, 1958.

CALCULATION OF SAMPLING ERRORS FOR INDEX NUMBERS

By K. S. BANERJEE

State Statistical Bureau, West Bengal

SUMMARY. In the construction of cost of living index (consumers' price index) numbers, the weight of a composite commodity is generally known in practice. The weights of all the constituent items of a composite commodity are not usually provided by a family budget inquiry. Criteria have been developed showing how unbiased estimate could be prepared of the index based on the knowledge of the weight of the composite commodity as a whole, and also how over-all precision of the estimated index could be increased and errors calculated.

1. INTRODUCTION

Calculation of sampling errors for Cost of Living Index (Consumers' Price Index) numbers under different sampling procedures has been illustrated in this note. The incidental observations, as made in this context, are expected to be helpful in increasing precision of the estimated index.

2. LASPEYRES' FORMULA

Laspeyres' aggregative formula, when expressed in full, takes the form,

$$100 \sum_i \sum_j p^1_{ij}\, q^0_{ij} / \sum_i \sum_j p^0_{ij}\, q^0_{ij}, \qquad \ldots \quad (2.1)$$

where p_{ij} and q_{ij} denote respectively the price and the quantity consumed of the j-th constituent item of the i-th composite commodity, and the upper subscripts, 1 and 0, denote respectively the period compared and the period of comparison. Formula (2.1) is reducible to the form,

$$\sum_i \sum_j w_{ij}\, r_{ij}, \qquad \ldots \quad (2.2)$$

where $r_{ij} = 100\, p^1_{ij}/p^0_{ij}$, $w_{ij} = p^0_{ij}\, q^0_{ij} / \sum_i \sum_j p^0_{ij}\, q^0_{ij}$ and $\sum_i \sum_j w_{ij} = 1$.

Formula (2.2) may further be reduced to the form,

$$\sum_i W_i r_i, \qquad \ldots \quad (2.3)$$

where r_i is the price relative of the i-th composite commodity (if it could be found), W_i, its weight, and $\sum_i W_i = 1$.

3. WEIGHTS OF CONSTITUENT ITEMS OF COMPOSITE COMMODITIES USUALLY NOT KNOWN IN PRACTICE

The weights are usually determined by a family budget survey. The questionnaire of such a budget inquiry usually provides for entries of expenditure on composite commodities. Weights of composite commodities only can thus be computed. As it is not possible, for practical reasons, to know of the expenditure on the constituent items of a composite commodity individually, the weights for the constituent items remain unknown. Had these weights been known, formula (2.2) could straight away be adopted for the construction of the index without making any error. But, as these weights remain usually unknown, a formula has to be preferred for which knowledge of the weight of a composite commodity as a whole should be sufficient. Formula (2.3) satisfies this condition. Only the weight of a composite commodity is involved in it.

4. CONSIDERATIONS ON WHICH PRECISION OF THE ESTIMATED INDEX DEPENDS

Precision of the calculated index will depend upon how a composite commodity is composed. Indeed, much will depend upon the manner in which the consumption items are grouped into composite commodities.

A group of items (composite commodity) may consist of a large number (finite) of constituent sub-items with each of which will be associated a price relative. As it would be impracticable to include all the constituent sub-items of a composite commodity in the construction, the construction has, of necessity, to be confined to a sample of the constituent items. The next question would then be to find out how best the sample could be drawn.

5. UNBIASED ESTIMATE OF THE INDEX AND ITS ERROR

In order to standardise the method of construction of index numbers, an unbiased estimate of the index was developed (1956a) under a simplifying assumption. The assumtion was that the composite commodities be so chosen that the expenditures on all the constituent single items of a composite commodity are the same. This assumption meant that the component items in commodity classifications were of equal importance in consumption. Under this assumption, the estimate for the index was obtained as

$$\sum_{i} \sum_{k=1}^{ni} \frac{W_i \, r_{ik}}{n_i}, \qquad \ldots \quad (5.1)$$

where r_{ik} is a price relative of the k-th constituent item of the i-th composite commodity. W_i, the weight of the i-th composite commodity, n_i, the number of items selected at random from N_i constituent items of the composite commodity, and $\sum_i W_i = 1$.

The variance of the estimate (5.1) was obtained as

$$\sum_{i} \frac{N_i - n_i}{N_i - 1} \; \frac{W_i^2 \, \sigma_{r_i}^2}{n_i}, \qquad \ldots \quad (5.2)$$

where $\sigma_{r_i}^2$ is the variance of r_{ik}.

It will be noticed that the weight of the composite commodity as a whole is only involved in the estimate.

The above assumption could be further generalised. The generalised assumption would be that the correlation coefficient, ρ_i, between the weights, w_{ij}, and the price relatives, r_{ij}, is '0'. The assumption of $\rho_i = 0$ includes the special cases of (i) equality of weights (expenditures), (ii) equality of price relatives, and (iii) equality of both weights and price relatives.

In order that the assumption be fulfilled, or nearly fulfilled, a commodity may have to be split into two or more composite commodities.

The degree of precision of the estimated index, as will be seen later, will depend upon the extent to which it has been possible to have the assumption realised in practice.

Under the generalised assumption also, the estimate of index will be the same as (5.1) and that of its variance the same as (5.2).

6. ESTIMATES UNDER DIFFERENT PROCEDURES AND THEIR ERRORS

Practices differ with regard to the treatment of composite commodities. One practice appears to be this : An arithmetic mean of prices of the sampled constituent items is drawn for the base period as well as for the period of comparison. The relative of such average prices is calculated, and this relative is then averaged with the weight. It has been shown (1956c) that the assumption underlying this practice is the equality of quantities consumed of the constituent sub-items of the composite commodity. Under this assumption, the estimate of the index gets approximated to

$$\sum_i \frac{\sum_k p^1_{ik}/n_i}{\sum_k p^0_{ik}/n_i} W_i, \qquad \dots \ (6.1)$$

where n_i is the number of constituent items selected from N_i component items of the composite commodity.

The variance for this estimate may be approximated to the same form as (5.2).

In this case also, the assumption of equality of the quantities may be generalised. The generalised assumption will be that the correlation between the price and the quantity consumed of the constituent sub-items of the composite commodity should be zero. The form of the estimated index and that of its variance will remain the same even when the assumption is thus generalised.

The average price of a composite commodity could be worked out (1956c) in the form of a harmonic mean of prices as well, and that, in a given situation. In that case also, the estimate of the index would be obtained in a form similar to that in (6.1), where the arithmetic mean would be replaced by the harmonic mean of prices. The variance for this estimate also coluld be approximated to the same form as given in (5.2).

Another estimate for the index of the i-th composite commodity, although not unbiased, could be obtained in the form[1]

$$\frac{\sum_k w_{ik}\, r_{ik}}{\sum_k w_{ik}}. \qquad \dots \ (6.2)$$

For this estimate, prior knowledge of the weights of the constituent items of the composite commodity would be essential, and this is a limitation. But, for this estimate also, the variance could be approximated to the same form as given in (5.2).

Again, an unbiased estimate of the index for the i-th composite commodity could as well be obtained in the form,

$$\frac{N_i}{n_i} \sum_k w_{ik}\, r_{ik}, \qquad \dots \ (6.3)$$

where $(w_{ik}\, r_{ik})$ is treated as a random variable, n_i of the N_i products of $(w_{ij}\, r_{ij})$ are selected for the sample, and where the weights, w_{ij}, of the constituent items of the composite commodities are assumed as known. The variance for this estimate is

$$\sum_i \frac{N_i - n_i}{N_i - 1} \frac{N_i^2\, \sigma^2_{wr}}{n_i}, \qquad \dots \ (6.4)$$

where σ^2_{wr} is the variance of $(w_{ij}\, r_{ij})$ in each i.

---

[1] Suggested recently by Irma Adelman (1958) : A new approach to the construction of index numbers, *The Review of Economics and Statistics*, Vol. XL No. 3, 240–249.

When, however, the weights of the constituent items of all the composite commodities are assumed as known, a different sampling procedure could be adopted. The procedure would be to select a sample of the constituent items with probabilities proportional to the weights of the constituent items (1958). If, under this sampling procedure, the sample is drawn *with replacement*, the estimate of the index of the i-th composite commodity will be

$$\frac{1}{n_i}\sum_{k=1}^{n_i} r_{ik} \qquad \ldots \quad (6.5)$$

While *sampling with replacement*, in the given context of selecting consumption items for price collection, may not be considered as ideal, *sampling without replacement* under such sampling procedure has the disadvantage of being a bit too complicated to be recommended for adoption into routine practice.

Inspite of the limitation coming from sampling *with replacement*, formula (6.3) could be used, if the weights of the constituent items were known. As stated before, such weights cannot be provided by a family budget inquiry, but good approximation to such weights may be obtained from market data (1956b, 1958) of prices and quantities sold of the constituent items.

It is interesting to note, however, that under different sampling procedures indicated in sections 5 as also under the first three procedures indicated in section 6, the variance of the estimate is given by (5.2).

7. AGREEMENT BETWEEN THE ESTIMATED INDEX AND THE TRUE VALUE[1] OF THE INDEX

In order to make possible the derivation of an unbiased estimate, the entire consumption will have to be divided into a number of homogeneous groups of commodities (composite commodities) such that the correlation (or covariance) between the weights and the price relatives of the constituent items of each composite commodity is zero. It has also been mentioned that the degree of precision of the estimate would depend upon the extent to which it has been possible to attain this objective. However good may be the care taken in this regard, some amount of covariance will exist although perhaps it may be negligibly small in respect of most of the commodities. A hypothetical example has been constructed with food items to illustrate this point. The illustration, though hypothetical, may be nearly representative of a genuine situation.

The corrected sum of products of the weights and the price relatives of the constituent items in respect of each of the 29 composite commodities, into which Food has been divided, has been indicated in Col. (8) of Table 1. It will be seen that for some composite commodities such as Rice (fine), potato, and fish (I), the corrected sum of products has been as large as +6.1645, +9.7000 and −5.7184 respectively. For these composite commodities, therefore, the condition of independence could not be secured to the extent it has been possible to do so in respect of the other composite commodities. The departure of the estimated index from the true value of the index (represented by the raw sum of products of the weights and the relatives) will, therefore, be large in respect of these commodities. In such cases, it will generally be observed that one or two (or more) of the constituent items stand apart from the rest of the constituent items to merit independent treatment. When such items are given an independent treatment, agreement between the estimated index and the true value of index will be closer.

---

[1] "True value" of the index is differentiated from the *true index* which has a meaning of its own in the literature on index numbers.

It is, therefore, clear that in respect of some composite commodities ρ_i may not be equal to zero, or in other words, $\left(\sum_j w_{ij}\, r_{ij} - W_i \frac{\Sigma\, r_{ij}}{N_i} \right)$ may not be equal to an absolute zero. This uncontrollable difference will be equal to $(N_i \rho_i \sigma_{wi} \sigma_{ri})$ which may be treated as *bias*. Owing to this bias, the root-mean-square variance of the estimated index for the i-th composite commodity will be equal to

$$\frac{N_i - n_i}{N_i - 1}\, \frac{W_i^2\, \sigma_{ri}^2}{n_i} + \Big[\; N_i\, \rho_i\, \sigma_{wi}\, \sigma_{ri} \;\Big]^2 \qquad \ldots \quad (7.1)$$

8. HOW CLOSER AGREEMENT MAY BE EFFECTED THROUGH A COMBINATION OF TWO OR MORE COMPOSITE COMMODITIES

Close agreement is feasible not only by a suitable subdivision of a composite commodity, but may be feasible also through a combination of two composite commodities. For instance, if the composite commodity (13) and the composite commodity (14) are combined, the resultant estimated index will be 594.00 as against 341.70+254.40 = 596.10. The difference is, therefore, 2.02 whereas the sum of the individual differences is 4.02. The resultant situation, therefore, gives rise to an increased precision of 2.00. The criterion giving rise to such increased precision may be worked out as follows :

Let there be two composite commodities (2.1) and (2.2), and let the indexes for the two be $\sum_j w_{1j} r_{1j}$ and $\sum_j w_{2j} r_{2j}$ respectively. Then, we have

$$\sum_j w_{1j}\, r_{1j} - \frac{W_1 \sum_j r_{1j}}{N_1} = N_1 \rho_1\, \sigma_{w_1}\, \sigma_{r_1}, \qquad \ldots \quad (8.1)$$

$$\sum_j w_{2j}\, r_{2j} - \frac{W_2 \sum_j r_{2j}}{N_2} = N_2 \rho_2\, \sigma_{w_2}\, \sigma_{r_2}, \qquad \ldots \quad (8.2)$$

$$\sum_j w_{1j}\, r_{1j} + \sum_j w_{2j}\, r_{2j} - \frac{W \left(\sum_j r_{1j} + \sum_j r_{2j}\right)}{N} = N\, \rho\, \sigma_w\, \sigma_s, \qquad \ldots \quad (8.3)$$

where $\sum_j w_{1j} = W_1$, $\sum_j w_{2j} = W_2$, $W_1 + W_2 = W$, $N_1 + N_2 = N$, and the other symbols have the usual significance. It can be easily shown that

$$N\, \rho\, \sigma_w\, \sigma_r = N_1 \rho_1 \sigma_{w1} \sigma_{r1} + N_2 \rho_2 \sigma_{w2} \sigma_{r2} + \frac{N_1 N_2 (\overline{W}_1 - \overline{W}_2)(\bar{r}_1 - \bar{r}_2)}{N} = \bar{\bar{W}} + \bar{\bar{B}}, \qquad \ldots \quad (8.3)$$

where $\overline{W}_1 = \frac{1}{N_1} \sum_j w_{1j}$, $\overline{W}_2 = \frac{1}{N_2} \sum_j w_{2j}$, $\bar{r}_1 = \frac{1}{N_1} \sum_j r_{1j}$, $\bar{r}_2 = \frac{1}{N_2} \sum_j r_{2j}$,

$\bar{\bar{W}} = N_1\, \rho_1\, \sigma_{w1}\, \sigma_{r1} + N_2\, \rho_2\, \sigma_{w2}\, \sigma_{r2}$, and $\bar{\bar{B}} = \frac{N_1 N_2}{N} (\overline{W}_1 - \overline{W}_2)\, (\bar{r}_1 - \bar{r}_2)$.

(8.3) means, as is well known, that the total sum of products (corrected) is equal to the sum of the sum of products within groups and the sum of products between groups, From this, it can easily be deduced that there will be perfect agreement between the estimated index and the true value of the index, if $\overline{\overline{W}}/B = -1$. There will be an improvement in precision upto the ratio, $\overline{\overline{W}}/\overline{\overline{B}}$, going down to -2.

Let us next turn to examine how the variance of the estimate may be affected when two composite commodities are combined into one. When the composite commodities (1) (2) are treated independently, the added root-mean square variance will be

$$\frac{N_1-n_1}{N_1-1}\,\frac{W_1^2\,\sigma^2_{r_1}}{n_1}+\frac{N_2-n_2}{N_2-1}\,\frac{W_2^2\,\sigma^2_{r_2}}{n_2}+\{[N_1\,\rho_1\,\sigma_{w_1}\,\sigma_{r_1}\,]^2+[N_2\rho_2\,\sigma_{w_2}\,\sigma_{r_2}\,]^2\}. \quad \ldots \quad (8.4)$$

But, when the two composite commodities are combined, and the sample size is kept at (n_1+n_2), the root-mean square variance of the estimated index for the combined composite commodity becomes

$$\frac{N-n}{N-1}\,\frac{W^2\,\sigma_r^2}{n}+[\,N_1\rho_1\,\sigma_{w_1}\,\sigma_{r_1}+N_2\rho_2\sigma_{w_2}\,\sigma_{r_2}+\frac{N_1N_2}{N}\,(\overline{W}_1-\overline{W}_2)(\bar{r}_1-\bar{r}_2)]^2, \quad \ldots \quad (8.5)$$

where

$$n = n_1+n_2,\text{ and } \sigma_r^2 = \frac{1}{N}\,(N_1\,\sigma^2_{r_1}+N_2\,\sigma^2_{r_2})+\frac{N_1N_2(\bar{r}_1-\bar{r}_2)^2}{N^2}.$$

Comparison of (8.4) with (8.5) will reveal whether there has been any increase in variance owing to the combination. If $\overline{\overline{W}}/\overline{\overline{B}} = -1$, the expression within the square bracket of (8.5) will be zero, and then there will no contribution owing to *bias*.

A word of precaution is necessary here. All necessary steps have to be taken against non-judicious grouping. It has already been indicated how non-judicious grouping may lead to a serious error (1959).

9. ERROR OF THE ESTIMATED INDEX CALCULATED UNDER DIFFERENT SAMPLING PROCEDURES AND THE ERRORS COMPARED

Table 1 shows, for each of the 29 composite commodities of food, the number of constituent items of each composite commodity (Col. 3), the total weight (Col. 4), the average of the price relatives (Col. 5), the true value of the index (Col. 6) i.e. the index calculated on all the constituent weights, and the index (Col. 7) calculated on the assumption that the correlation (or co-variance) between the weights and the price relatives of the constituent items is zero. The departure of the estimated index from the true value of the index is shown in Col. 8.

The difference in Col. (8) represents the bias (or what may be called the uncontrollable error), and is positive for some commodities and negative for others. The sum of all these differences is +0.1184 representing the over-all departure from the true value of the index of Food, which is 135.84. From the true value of the index the difference of the estimated index, 135.72, may be considered to be quite small for all practical purposes.

The correlation coefficient, ρ_i, and the square-roots of the variances of the weights and the relatives of the constituent items have been calculated and incorporated in Table 1 to show the contribution of each of these factors to the difference in Col. (8).

If, in respect of any composite commodity, the value of this difference is large, it would call for a closer scrutiny of the situation which might lead to a further sub-division of the composite commodity.

Table 2 shows the estimate for each composite commodity prepared from a sample of consumption items drawn at random from the constituent items of a composite commodity. The size of the sample has been indicated in Col. (4). The sample size has been fixed in such a manner that it is nowhere less than 3 and nowhere larger than 10. In case the number of constituent items of a constituent commodity is 3 or less, all the constituent items have been taken into account. The size of the total sample is thus 113 drawn from the total of 284 items.

The estimated index is shown in Col. (8). Col. (9) shows the departure of the estimated index from the true value of the index . These departures, as before, are somewhere positive and somewhere negative. Their sum is +0.0562, which is the over-all departure of the estimated index from the true value of the index.

The variance for the estimated index is indicated in Col. (11). The sum of the variances is 0.1610, to which the total of Col. (12) of Table 1 (i.e. the total of the squares of the biases) may be added to find the root-mean-square variance of the estimated index. The contribution owing to the bias is 0.0236 being less than one-sixth the value of the total variance. The root-mean-square error is 0.4296.

In Table 2 the size of the sample for each commodity was chosen rather arbitrarily. If, however, the size of the sample were chosen in accordance with the principle of optimum allocation as outlined before (1956a), the total variance for the estimated index would be much less. This has been illustrated in Table 3.

The total size has been taken as 115 which is of the same order as 113, the size of the sample illustrated in Table 2. The distribution of this total size by optimum allocation (that is, by an allocation which is proportional to ($\sigma_{ri} W_i$) is indicated in Col. (3) of Table 3. It was noticed that in respect of a few commodities, the worked out size came out to be more than the number of constituent items of the composite commodity concerned. The sample size for such composite commodities was, however, restricted to the total number of the constituent items. In respect of some composite commodities, again, the worked out size came out as less than 3. For such commodities, the sample size has been fixed at 3. These restrictions increased the total sample size to 118.

It will be noticed in Table 3 that the estimated index has come out to be 135.69 which is less than the true value of the index by only 0.15. The variance of the estimate is 0.0266 which is roughly one-sixth the variance obtained for the sample illustrated in Table 3. This shows that optimum allocation will be highly potential in increasing the precision of the estimate. The root-mean-square error is 0.2241.

It has already been indicated that sampling of constituent items with probability proportional to the constituent weights, where prior knowledge of the constitent weights exists would also lead to an unbiased estimate. A sample *with replacement* was taken of the constituent items with probability proportional to the constituent weights. The calculations have been indicated in Table 4.

TABLE 1. DIFFERENCE OF THE ESTIMATED INDEX (UNDER THE STATED ASSUMPTION) FROM THE TRUE VALUE WHEN ALL THE CONSTITUENT ITEMS OF A COMPOSITE COMMODITY ARE INCLUDED IN THE CONSTRUCTION

serial no.	items of food	N_i	W_i	$\Sigma r_{ij}/N_i$	$\Sigma w_{ij}.r_{ij}$	$\Sigma r_{ij}/N_i \times W_i$	col. (6)– col. (7)	"ρ_i"	$\overset{*}{\sigma}_{w_i}$	$\overset{*}{\sigma}_{r_i}$	(bias)^2 col (8)^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1.	rice (fine)	7	5.00	129.8571	655.45	649.2855	6.1645	0.3907	0.3583	6.2906	37.9986
2.	rice (medium)	7	10.00	119.8571	1195.80	1198.5710	−2.7710	−0.1288	0.7166	4.2905	7.6807
3.	rice (coarse)	7	15.00	122.7143	1843.65	1840.7145	2.9355	0.2342	1.0748	1.6660	8.6172
4.	rice product-I	2	2.00	120.0000	240.50	240.0000	0.5000	1.0000	0.2500	1.0000	0.2500
5.	rice product-II	2	0.50	117.5000	58.70	58.7500	−0.0500	−1.0000	0.0500	0.5000	0.0025
6.	rice product-III	2	0.25	117.0000	29.20	29.2500	−0.0500	−0.9434	0.0265	1.0000	0.0025
7.	wheat	5	4.25	104.4000	444.85	443.7000	1.1500	0.3019	0.2429	3.1368	1.3225
8.	barley	4	0.15	101.2500	15.17	15.1875	−0.0175	−0.5271	0.0100	0.8292	0.0003
9.	tea	8	1.75	132.0000	228.55	231.0000	−2.4500	−0.3512	0.1145	7.6158	6.0025
10.	pulses	8	3.50	178.3750	626.30	624.3125	1.9875	0.5013	0.2655	1.8667	3.9502
11.	mustard oil	8	6.00	101.7500	610.35	610.5000	−0.1500	−0.0137	0.4507	3.0311	0.0225
12.	ghee	6	0.50	113.0176	55.25	56.5088	−1.2588	−0.7097	0.0374	7.9039	1.5846
13.	hydrogenated vegetable oil	6	2.50	135.8333	341.70	339.5833	2.1167	0.3084	0.2592	4.4127	4.4804
14.	sugar	4	2.00	126.2500	254.40	252.5000	1.9000	0.3257	0.3374	4.3229	3.6100
15.	molasses	4	2.50	189.5000	475.40	473.7500	1.6500	0.3480	0.3897	3.0414	2.7225
16.	milk	4	3.00	91.5000	274.50	274.5000	0.0000	0.0000	0.2500	1.5000	0.0000
17.	butter	3	0.15	134.0000	19.65	20.1000	−0.4500	−0.9932	0.0245	6.1644	0.2025
18.	curd	3	0.07	130.6667	9.15	9.1467	0.0033	0.0000	0.0000	0.4713	0.0000
19.	sweetmeats	3	0.03	129.3333	3.88	3.8800	0.0000	0.0000	0.0000	1.2473	0.0000
20.	potato	6	4.00	111.0000	453.70	444.0000	9.7000	0.7783	0.6681	3.1091	94.0900
21.	onion	2	1.00	111.5000	111.75	111.5000	0.2500	1.0000	0.2500	0.5000	0.0625
22.	spices and salt	13	4.90	376.6154	1846.40	1845.4155	0.9845	0.1050	0.3384	2.1318	0.9692
23.	meat	4	1.00	120.2500	123.80	120.2500	3.5500	0.8250	0.1697	6.3394	12.6025
24.	fish I	43	8.03	143.2800	1144.82	1150.5384	−5.7184	−0.2514	0.2032	2.6033	32.7001
25.	fish II	3	3.25	144.0000	467.99	468.0000	−0.0100	−0.0000	0.0000	2.1603	0.0001
26.	egg	3	1.50	126.0000	187.25	189.0000	−1.7500	−0.9704	0.2042	2.9439	3.0625
27.	nonleafy vegetables	45	8.47	105.5778	891.44	894.2440	−2.8040	−0.1115	0.2581	2.1653	7.8624
28.	leafy vegetables	41	4.90	105.6341	516.08	517.6071	−1.5271	−0.0652	0.2615	2.1837	2.3320
29.	fruits	31	3.80	121.2581	458.74	460.7808	−2.0408	−0.1240	0.2309	2.3551	4.3714
	total	284	100.00	133.8028	13584.42	13572.5756	11.8444	0.0053			236.5022

*Relates to the finite population of size given in col (3).

TABLE 2. DIFFERENCE OF THE ESTIMATED INDEX FROM THE TRUE VALUE WHEN THE INDEX IS ESTIMATED FROM A SAMPLE OF CONSTITUENT ITEMS OF A COMPOSITE COMMODITY (VARIANCE OF THE ESTIMATED INDEX ALSO SHOWN)

serial no.	items of food	N_i	n_i	W_i	$\Sigma r_{ik}/n_i$	$\Sigma w_{ij} r_{ij}$	$\frac{\Sigma r_{ik}}{n_i} . W_i$	col. 7-col. 8	$\sigma^{2*}_{r_i}$	$\frac{N_i-n_i}{N_i-1} \times \frac{W_i^2 \sigma_{r_i}^2}{n_i}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1.	rice (fine)	7	3	5.00	131.0000	655.4500	655.0000	0.4500	36.0000	200.0000
2.	rice (medium)	7	3	10.00	120.3333	1195.8000	1203.3330	−7.5330	49.3334	1096.3000
3.	rice (coarse)	7	3	15.00	123.0000	1843.6500	1845.0000	−1.3500	1.0000	49.9950
4.	rice product-I	2	2	2.00	120.0000	240.5000	240.0000	0.5000	1.0000	—
5.	rice product-II	2	2	0.50	171.5000	58.7000	58.7500	−0.0500	0.2500	—
6.	rice product-III	2	2	0.25	117.0000	29.2000	29.2500	−0.0500	1.0000	—
7.	wheat	5	3	4.25	103.6667	444.8500	440.5835	4.2665	22.3334	67.2322
8.	bread	4	3	0.15	101.0000	15.1700	15.1500	0.0200	1.0000	0.0025
9.	tea	8	3	1.75	129.0000	228.5500	225.7500	2.8000	16.0000	11.6666
10.	pulses	8	3	3.50	179.0000	626.3000	626.5000	−0.2000	4.0000	11.6706
11.	mustard oil	8	3	6.00	100.0000	610.3500	600.0000	10.3500	1.0000	8.5716
12.	ghee	6	3	0.50	111.0000	55.2500	55.5000	−0.2500	91.0000	4.5500
13.	hydrogenated vegetable oil	6	3	2.50	137.6667	341.7000	344.1668	−2.4668	9.3334	11.6669
14.	sugar	4	3	2.00	128.6667	254.4000	257.3334	−2.9334	2.3334	1.0372
15.	molasses	4	3	2.50	188.6667	475.4000	471.6668	3.7332	14.3334	9.9538
16.	milk	4	3	3.00	92.0000	274.5000	276.0000	−1.5000	3.0000	2.9997
17.	butter	3	3	0.15	134.0000	19.6500	20.1000	−0.4500	38.0000	—
18.	curd	3	3	0.07	130.6667	9.1500	9.1467	0.0033	0.2222	—
19.	sweetmeats	3	3	0.03	129.3333	3.8800	3.8800	0.0000	1.5554	—
20.	potato	6	3	4.00	113.3333	453.7000	453.3332	0.3668	10.3334	33.0672
21.	onion	2	2	1.00	111.5000	111.7500	111.5000	0.2500	0.2500	—
22.	spices and salt	13	5	4.90	376.2000	1846.4000	1843.3800	3.0200	9.2000	29.4531
23.	meat	4	3	1.00	118.6667	123.8000	118.6667	5.1333	60.3333	6.7037
24.	fish-I	43	10	8.03	142.8000	1144.8200	1146.6840	−1.8640	3.9556	20.0084
25.	fish-II	3	3	3.25	144.0000	467.9900	468.0000	−0.0100	4.6667	—
26.	egg	3	3	1.50	126.0000	187.2500	189.0000	−1.7500	8.6667	—
27.	nonleafy vegetables	45	10	8.47	105.2000	891.4400	891.0440	0.3960	5.2889	30.1814
28.	leafy vegetables	41	10	4.90	106.1000	516.0800	519.8900	−3.8100	4.7667	8.8693
29.	fruits	31	10	3.80	121.1000	458.7400	460.1800	−1.4400	6.5444	6.6150
	total	284	113	100.00	134.4071	13584.4200	13578.7881	5.6319		1610.5442

*Estimated from the sample in case of sampling.

TABLE 3. DIFFERENCE OF THE ESTIMATED INDEX FROM THE TRUE VALUE WHEN THE INDEX IS ESTIMATED FROM A SAMPLE OF CONSTITUENT ITEMS OF A COMPOSITE COMMODITY UNDER OPTIMUM ALLOCATION (VARIANCE OF THE ESTIMATED INDEX ALSO SHOWN)

serial no.	items of food	N_i	n_i	W_i	$\Sigma\, r_{ik}/n_i$	$\Sigma\, w_{ij}\, r_{ij}$	$\frac{\Sigma\, r_{ik}}{n_i}.\ W_i$	Difference col. 7-col. 8	ϕ^{2*}_{ri}	$\frac{N_i-n_i}{N_i-1}\cdot\frac{W^2\sigma^{2'}_{ri}}{n_i}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1.	rice (fine)	7	7	5.00	129.8571	655.4500	649.2855	6.1645	39.5714	—
2.	rice (medium)	7	7	10.00	119.8571	1195.8000	1198.5710	−2.7710	18.4082	—
3.	rice (coarse)	7	7	15.00	122.7143	1843.6500	1840.7145	2.9355	3.2381	—
4.	rice product-I	2	2	2.00	120.0000	240.5000	240.0000	0.5000	1.0000	—
5.	rice product-II	2	2	0.50	117.5000	58.7000	58.7500	−0.0500	0.2500	—
6.	rice product-III	2	2	0.25	117.0000	29.2000	29.2500	−0.0500	1.0000	—
7.	wheat	5	3	4.25	105.0000	444.8500	446.2500	−1.4000	21.0000	63.2187
8.	bread	4	3	0.15	101.0000	15.1700	15.1500	0.0200	1.0000	0.0025
9.	tea	8	5	1.75	129.4000	228.5500	226.4500	2.1000	12.8000	3.6004
10.	pulses	8	3	3.50	176.6666	626.3000	618.3331	7.9669	2.3334	6.8057
11.	mustard oil	8	7	6.00	101.0000	610.3500	606.0000	4.3500	7.0000	5.1444
12.	ghee	6	3	0.50	113.0000	55.2500	56.5000	−1.2500	133.0000	6.6500
13.	hydrogenated vegetable oil	6	3	2.50	135.0000	341.7000	337.5000	4.2000	52.0000	65.0000
14.	sugar	4	3	2.00	125.3333	254.4000	250.6666	3.7334	32.3334	14.3689
15.	molasses	4	3	2.50	190.6666	475.4000	476.6665	−1.2665	10.3334	7.1752
16.	milk	4	3	3.00	91.5500	274.5000	274.5000	0.0000	3.0000	2.9997
17.	butter	3	3	0.15	134.0000	19.6500	20.1000	−0.4500	38.0000	—
18.	curd	3	3	0.07	130.6666	9.1500	9.1467	0.0033	0.2222	—
19.	sweetmeats	3	3	0.03	129.3333	3.8800	3.8800	0.0000	1.5554	—
20.	potato	6	5	4.00	111.4000	453.7000	445.6000	8.1000	13.3000	8.5120
21.	onion	2	2	1.00	111.5000	111.7500	111.5000	0.2500	0.2500	—
22.	spices and salt	13	4	4.90	377.0000	1846.4000	1847.3000	−0.9000	2.0000	9.0038
23.	meat	4	3	1.00	123.0000	123.8000	123.0000	0.8000	7.0000	0.7777
24.	fish-I	43	11	8.03	143.5455	1144.8200	1152.6704	−7.8504	10.8656	48.5278
25.	fish-II	3	3	3.25	144.0000	467.9900	468.0000	−0.0100	4.6667	—
26.	egg	3	3	1.50	126.0000	187.2500	189.0000	−1.7500	8.6667	—
27.	non-leafy vegetables	45	7	8.47	106.7143	891.4400	903.8701	−12.4301	2.9030	2.5692
28.	leafy vegetables	41	4	4.90	104.0000	516.0800	509.6000	6.4800	2.0000	11.1046
29.	fruits	31	3	3.80	121.3333	458.7400	461.0665	−2.3265	2.3334	10.4825
	total	284	117	100.00	133.6581	13584.4200	13569.3209	15.0991		265.9431

*Estimated from the sample in case of sampling.

TABLE 4. DIFFERENCE OF THE ESTIMATED INDEX FROM THE TRUE VALUE WHEN THE INDEX IS ESTIMATED FROM A SAMPLE OF CONSTITUENT ITEMS OF A COMPOSITE COMMODITY, THE SAMPLE BEING DRAWN, WITH REPLACEMENT, WITH PROBABILITY PROPORTIONAL TO THE WEIGHT. (VARIANCE OF ESTIMATED INDEX ALSO SHOWN)

serial no.	items of food	N_i	n_i'	W_i	$\Sigma\, r_{ik}/n_i'$	$\Sigma\, r_{ij}.\, w_{ij}$	$\frac{\Sigma\, r_{ik}}{n_i'}.\, W_i$	col. (7)-col. (8)	$\sigma^{2*}_{r_{ik}}$	$W^2 .\ \sigma^2_{r_{ik}/n_i'}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1.	rice (fine)	7	3	5.00	129.0000	655.4500	645.0000	10.4500	12.0000	100.0000
2.	rice (medium)	7	3	10.00	119.0000	1195.8000	1190.0000	5.8000	52.0000	1733.3300
3.	rice (coarse)	7	3	15.00	123.0000	1843.0000	1845.0000	–1.3500	1.0000	74.9925
4.	rice product-I	2	2	2.00	120.0000	240.5000	240.0000	.5000	1.0000	2.0000
5.	rice product-II	2	2	.50	117.0000	58.7000	58.7500	– .0500	0.2500	.0312
6.	rice product-III	2	2	.25	117.0000	29.2000	29.2500	– .0500	1.0000	.0312
7.	wheat	5	3	4.25	102.3333	444.8500	434.9165	9.9335	6.3334	38.1317
8.	bread	4	3	.15	101.0000	15.1700	15.1500	.0200	1.0000	.0075
9.	tea	8	3	1.75	128.3333	228.5500	224.5833	3.9667	1.3334	1.3613
10.	pulses	8	3	3.50	179.3333	626.3000	627.6666	–1.3666	1.3334	5.4451
11.	mustard oil	8	3	6.00	99.6666	610.3500	598.0000	12.3500	21.3334	255.9996
12.	ghee	6	3	.50	114.6666	55.2500	57.3333	–2.0833	86.3334	7.1945
13.	hydrogenated vegetable oil	6	3	2.50	133.6666	341.7000	334.1665	7.5335	37.3334	77.7781
14.	sugar	4	3	2.00	125.0000	254.4000	250.0000	4.4000	28.0000	37.3332
15.	molasses	4	3	2.50	192.3333	475.4000	480.8332	–5.4332	.3334	.6944
16.	milk	4	3	3.00	90.0000	274.5000	270.0000	4.5000	3.0000	9.0000
17.	butter	3	3	.15	134.0000	19.6500	20.1000	– .4500	38.0000	.2850
18.	curd	3	3	.07	130.6667	9.1500	9.1467	.0033	.2222	.0004
19.	sweetmeats	3	3	.03	129.3333	3.8800	3.8800	.0000	1.5554	0.0005
20.	potato	6	3	4.00	115.0000	453.7000	460.0000	–6.3000	12.0000	64.0000
21.	onion	2	2	1.00	111.5000	111.7500	111.5000	.2500	.2500	.0312
22.	spices and salts	13	5	4.90	376.8000	1846.4000	1846.3200	.0800	9.2000	4.4419
23.	meat	4	3	1.00	120.3333	123.8000	120.3333	3.4667	80.3334	26.7778
24.	fish-I	43	10	8.03	142.8000	1144.8200	1146.6840	–1.8640	3.9556	25.5080
25.	fish-II	3	3	3.25	144.0000	467.9900	468.0000	– .0100	4.6667	16.4310
26.	egg	3	3	1.50	126.0000	187.2500	189.0000	–1.7500	8.6667	6.5000
27.	non-leafy vegetables	45	10	8.47	104.9000	891.4400	888.5030	2.9370	3.6556	26.2285
28.	leafy vegetable	41	10	4.90	104.8000	516.0800	513.5200	2.5600	2.6222	6.2954
29.	fruits	31	10	3.80	120.7000	458.7400	458.6600	.0800	3.7889	5.4713
	total	284	113	100.00	134.2301	13584.4200	13536.2964	48.1236	—	2525.3013

*Estimated from the sample in case of sampling.

It will be observed that the estimated index has come out as 135.36, and its estimated variance as 0.2525. This variance is higher than that obtained in Table 2 and Table 3. As the estimate is unbiased, there is no contribution by "bias." The standard error is 0.5020.

The results may be summarised as follows :

sampling procedure as in	nature of the estimate	true value of the index	estimated index	variance	square of the bias	root-mean-square error or standard error
table 2	biased	135.84	135.79	.1610	0.0236	0.4296
table 3	biased (optimum allocation)	135.84	135.69	.0266	0.0236	0.2241
table 4	unbiased	135.84	135.36	.2525	×	0.5020

Fromula (6.3) which also affords an unbiased estimate was not used in the construction of the index, because formula (6.3) is not expected to be suitable for the type of data illustrated above. The estimated index will be wide away the true value with a large variance.

10. Remarks

The sampling procedure in Table 2 which affords the highest precision may be adopted into routine practice. Its main advantage is that the estimate can be prepared without a knowledge of weights of the constituent items.

11. Acknowledgment

Assistance in computation rendered by Shri A. Chowdhury is gratefully acknowledged.

References

Banerjee, K. S. (1956a) : A note on the optimal allocation of consumptive items in the construction of a cost of living index. *Econometrica*, **24**, No. 3, 294–295.

——— (1956b) : A comment on the construction of price index numbers. *Applied Statistics*, **5**, No. 3, 207–210.

——— (1956c) : A note in the treatment of composite items in the construction of cost of living index numbers, *Cal. Stat. Ass. Bull.*, **7**, No. 25, 35–40.

——— (1958) : Probability selection of the constituent items of a composite item in the construction of cost of living index numbers. *Cal. Stat. Ass. Bull.*, **8**, Nos. 30 & 31, 106–109.

Paper received : June, 1959.

TYPE-STUDY ON PEAK PERIOD IN HARVESTING *AMAN* PADDY, WEST BENGAL, DECEMBER 1954-JANUARY 1955

By R. K. SOM, G. C. BHATTACHARYYA *and* N. K. NAMBOODIRI
Indian Statistical Institute, Calcutta

SUMMARY. There is a general impression that there is much surplus population in rural areas of India: but as Mahalanobis (1955) has stated, this is still an open question, at least at the present level of technology. The results of a Type-study conducted in five villages in West Bengal in 1954-55 suggest that it may be more economical to maintain a supply of labour which would be required at the time of busy periods of agricultural activities, even if they remain comparatively under-employed in the remaining period of the year.

1. Introduction

1. *Objective*. The importance of basic studies in regard to mobilisation of manpower in rural areas in the peak periods of agricultural activities in the background of national planning was first pointed by P. C. Mahalanobis.[1] "There is a general impression that there is much surplus population in villages, but this is still an open question. It is possible that although the agricultural population is very large, it is needed to supply labour essential at a period of peak load (at the stage of transplanting of the rice crop, or at the stage of harvesting of rice and other crops). If one portion of the agricultural population is drawn away from the villages, there would be some decrease in the outturn of crops. If the decrease in the outturn of foodgrains due to such labour shortage is greater than the net contribution (i.e., total contribution minus consumption of food) of the transferred workers to the national income then obviously it would be economical, in a broad sense, to keep these labour households within the village to supply the peak load of labour required for full agricultural production, even if they do not have other gainful work during the rest of the year. On this view, there would be a biological balance between the size of the agricultural population and the volume of crop production, and it is conceivable that a decrease in agricultural population (without introducing labour-saving devices) might diminish the real income per person" (Mahalanobis, 1955).

1.2. This note presents the results of a type-study carried out in five villages in West Bengal, with a view to measuring the manpower utilisation in the peak week in *Aman* harvesting, December 1954 — January 1955. The villages were so selected as to represent, on a small scale, the varying geographical and occupational patterns obtaining in the State. The very nature of the study precluded the construction of "estimates" proper; regardless, the results could be useful in posing the relevant questions and in evolving the methodology for future studies.

---

[1] Later incorporated in (Mahalanobis, 1955, p. 10).

2. Village characteristics

2.1. *Village particulars.* The location, area, population and its distribution in the different livelihood classes in the five villages have been given in Tables 1 and 2.

TABLE 1. LOCATION, AREA AND POPULATION OF THE VILLAGE[1]

district	village	distance (miles) from: town	distance (miles) from: railway station	area (acres)	population per acre	number of occupied houses	total number of plots	percentage of paddy plots	percentage of area under paddy
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. 24-Parganas	*G*	10	4	508	1.58	184	1350	54.1	64.9
2. Burdwan	*N*	12	7	469	2.28	370	1400	45.6	73.1
3. West Dinajpur	*C*	22	37	230	0.35	19	276	59.1	87.4
4. Midnapore	*D*	15	24	1104	1.21	268	3566	64.3	80.6
5. Cooch Behar	*S*	3	3	545	0.96	115	2007	48.8	50.5

---

[1] The full name of the villages are: Uttar Gazipur (*G*), Nandanghat (*N*), Chandipara (*C*), Daudpur (*D*), and Saulmurichota (*S*).

TABLE 2. PERCENTAGE DISTRIBUTION OF POPULATION IN LIVELIHOOD CLASSES, CENSUS 1951

village	population	percentage of population in livelihood classes							
		agricultural				non-agricultural			
		cultivators of land wholly or mainly: owned	cultivators of land wholly or mainly: unowned	cultivating labourers	non-cultivating owners of land	production other than agricultural	commerce	transport	other services and miscellaneous sources
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. *G*	800	75.2	0.8	11.8	0.1	5.2	2.5	0.8	3.6
2. *N*	1069	3.4	15.1	12.1	0.7	8.8	18.3	6.0	35.6
3. *C*	80	78.8	21.2	—	—	—	—	—	—
4. *D*	1337	36.7	47.6	5.2	0.2	0.9	1.4	0.5	7.5
5. *S*	523	71.5	28.5	—	—	—	—	—	—

The descriptive location, area, and population of the villages are given below in Table 3. Villages near to, and remote from, towns and railway stations, with population densities low to high, and means of livelihood varying from fully agricultural to mainly non-agricultural, were covered in this study. The location of the villages has been shown in the map.

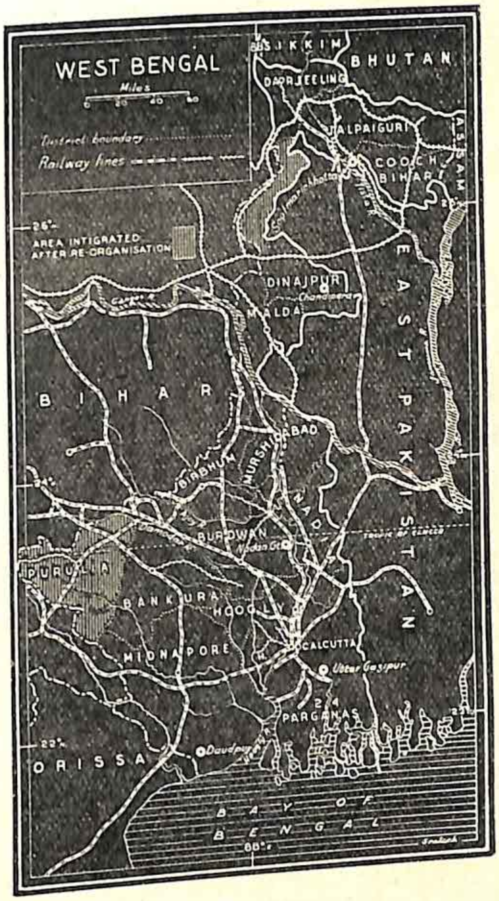


TABLE 3. DESCRIPTION OF LOCATION, AREA AND POPULATION OF THE VILLAGES

village (1)		district (2)	distance (3)	population density (4)	percentage of paddy plots (5)	area under paddy (6)	means of livelihood (7)
1.	*G*	24-Parganas	intermediate	medium	medium	medium	mainly agricultural
2.	*N*	Burdwan	intermediate	high	low	medium	mainly non-agricultural
3.	*C*	West Dinajpur	remote	low	medium	high	fully agricultural
4.	*D*	Midnapore	next remote	medium	high	high	mainly agricultural
5.	*S*	Cooch Behar	near	medium	low	low	fully agricultural

3. POPULATION CHARACTERISTICS

3.1. *Age distribution.* The proportion of population in the young age range 0–16 was maximum (49 percent) in village *D* (with a high proportion of area under paddy) and the proportion in the working age group 17–61 maximum (63 percent) in "non-agricultural" village *N*.

TABLE 4. PERCENTAGE DISTRIBUTION OF SAMPLE PERSONS BY AGE GROUP

age-group (years)	village				
	G	*N*	*C*	*D*	*S*
(1)	(2)	(3)	(4)	(5)	(6)
1. 0—16	43.1	35.9	38.9	49.0	36.3
2. 17—61	53.8	62.8	60.0	51.0	60.4
3. 62 & above	3.1	1.3	1.1	—	3.3
4. all ages	100.0	100.0	100.0	100.0	100.0
(number of sample persons)	(65)	(78)	(90)	(51)	(91)

3.2. *Normal activity status.* The activity status of a person describes his position in the economy as an employer, employee, own account worker, unpaid household enterprise worker (these four categories constituting the gainfully employed), as an unemployed seeking employment (this along with the gainfully employed constituting the labour force) or as a family member engaged in domestic work, student or in other non-productive activities. The percentage distribution of the sample persons by normal acitivity status is shown in Table 5. The maximum proportion of gainfully employed (50 percent) was observed in "non-agricultural" village *N* and the minimum (26 percent) in village *G*. It is interesting to note that the proportion returned as "too young" varied more than the proportion in the young age group 0–16; the ratio of persons with activity status "too young" to persons in age group 0–16 was maximum (0.84) in village *D* and minimum (0.39) in non-agricultural village and "near" village. This is indicative of the different social mores obtaining in the villages.

TABLE 5. PERCENTAGE DISTRIBUTION OF SAMPLE PERSONS BY NORMAL ACTIVITY STATUS

activity status	village				
	G	*N*	*C*	*D*	*S*
1. employer	—	5.1	—	—	—
2. employee	3.1	23.1	7.8	15.7	1.1
3. own account worker	23.1	11.5	21.1	11.8	14.3
4. unpaid household labour	—	10.3	3.3	—	22.0
5. gainfully employed	26.2	50.0	32.2	27.5	37.4
6. family members engaged in domestic work	30.8	28.2	32.3	21.5	31.8
7. student	9.2	5.1	3.3	2.0	12.1
8. rentiers, pensioners and remittance holders	—	1.3	—	5.9	—
9. beggars, persons living on charity and on ungainful activity	4.6	—	—	—	—
10. old and infirm	3.1	1.3	4.4	2.0	4.4
11. too young	26.1	14.1	27.8	41.1	14.3
12. total	100.0	100.0	100.0	100.0	100.0
(number of sample persons)	(65)	(78)	(90)	(31)	(91)

3.3. *Shift in activity status.* The shift in activity status of household members during the busy week from the normal activity status, though small, was mainly experienced by persons normally outside labour force — persons not essentially required in domestic work and students becoming unpaid household labour.

3.4. As a measure of immobility in activity status, i.e., the tendency to remain in the same activity status in the busy week as the normal activity status, Indices of Inertia were calculated. The index of overall Inertia is defined as the percentage of persons having the same activity status both normally and in the busy week as well. The Index of Inertia within the labour force is the percentage of persons normally in labour force, who had the same (normal) activity status in the busy week; the Index of Inertia outside the labour force is similarly defined. The Labour Force Entry Index is defined as the percentage of persons normally outside labour force who entered the labour force during the busy week. These indices for the five villages are shown in Table 6. The index most relevant for our purpose is the Labour Force Entry Index. In three villages, however, persons normally outside labour force entered the labour force in the busy week. In "remote—low population density" village *C* the high index of 32 for females is observed, occasioned by 15, out of a total 29, females not essentially required in domestic work who worked as unpaid household labour during the busy period.

TABLE 6. INDEX OF INERTIA IN ACTIVITY STATUS BY VILLAGE AND SEX OF PERSONS

village	overall inertia			inertia within labour force			inertia outside labour force			labour force entry index		
	male	female	total	male	female	total	male	female	total	male	female	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. *G*	100	100	100	100	—	100	100	100	100	—	—	—
2. *N*	98	100	99	97	—	79	100	100	100	—	—	—
3. *C*	98	67	81	100	100	100	100	97	98	7	32	26
4. *D*	88	100	94	85	100	86	100	100	100	8	—	3
5. *S*	92	100	98	94	—	94	100	100	100	13	—	4

4. Harvesting detail in peak week

4.1. *Location of peak week.* The peak week, which corresponded to the seven day for which the moving total of the daily harvested areas was maximum, started in the last week of December in all the five villages; no departure from the "normal" date of commencement was observed in the district as a whole (Table 7).

TABLE 7. DATE OF COMMENCEMENT OF HARVESTING IN THE DISTRICTS AND OF THE BUSY WEEK IN THE VILLAGES

district	date of commencement of *aman* harvesting 1954-55[1]	village	date of commencement of peak week, *aman* harvesting, 1954-55
(1)	(2)	(3)	(4)
1. 24-Parganas	1st week of December	*G*	22 December
2. Burdwan	1st week of December	*N*	20 December
3. West Dinajpur	1st week of December	*C*	26 December
4. Midnapore	end of November	*D*	19 December
5. Cooch Behar	middle of November	*S*	21 December

---

[1] *Supplement to Calcutta Gazette*, April 21, 1955, p. 383.

4.2. *Man-days spent on harvesting and related operations.* Of the man-days worked for harvesting in the peak week, the proportion spent directly on harvesting varied from 90.8 percent in village *G* to 100 percent in "near" village *S*. The subsidiary activities related to transport, storing, watching, supplying meals, etc.

TABLE 8. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED DURING THE PEAK WEEK OF HARVESTING BY PERSONS ENGAGED DIRECTLY IN HARVESTING AND PERSONS NOT ENGAGED IN HARVESTING

village	percentage of man-days worked			number of man-days
	directly engaged in harvesting	not engaged in harvesting	total	
(1)	(2)	(3)	(4)	(5)
1. *G*	90.8	9.2	100.0	303
2. *N*	95.7	4.3	100.0	649
3. *C*	97.5	2.5	100.0	436
4. *D*	97.9	2.1	100.0	573
5. *S*	100.0	—	100.0	1201

4.3. *Man-days spent on harvesting during the peak week by age and sex.* These are given in Table 9. As could be expected, a large majority of man-days was contributed by persons in the working age group 17–61; in the young age range 0–16 years, the maximum contribution (21 percent) was made in "remote" village *C*. The maximum proportion of man-days worked by females was also in village *C* (24 percent), "non-agricultural" village *N* coming next with 16 percent. No contribution was made by females in village *G* and in the other two villages the female contribution was very small.

TABLE 9. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED IN HARVESTING DURING THE PEAK WEEK BY AGE GROUP AND SEX

village	age groups (years)										number of man-days
	0–16			17–61			62–above	all			
	male	female	total	male	female	total	male	male	female	total	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1. *G*	4.7	—	4.7	95.3	—	95.3	—	100.0	—	100.0	275
2. *N*	2.6	10.8	13.4	81.8	4.8	86.6	—	84.4	15.6	100.0	621
3. *C*	16.5	4.7	21.2	59.7	19.1	78.8	—	76.2	23.8	100.0	425
4. *D*	10.0	1.2	11.2	81.4	6.3	87.7	1.1	92.5	7.5	100.0	561
5. *S*	3.7	—	3.7	93.9	2.3	96.2	0.1	97.7	2.3	100.0	1201

4.4. *Distribution of man-days by normal occupation.* The percentage distribution of man-days worked in harvesting during the peak week by normal occupation and sex has been given in Table 10 ; the agricultural occupations related to farming, cultivation, share-cropping and agricultural labour, but excluded animal husbandry, fishery and hunting. The participation of persons with normal non-agricultural occupations in the "non-agricultural" village *N* and of persons outside labour force in the "remote" village *C* and in village *D* (with a high proportion of paddy plots) is to be noticed in this Table. In "non-agricultural" village *N* the majority of man-days (54 percent) was contributed by persons with normal non-agricultural occupations. The comparatively high proportion (30 percent) of man-days was contributed by persons normally outside labour force in "remote—low population density" village *C*, while in village *D* (the "next remote" village), the proportion was 12.1 percent. The proportion of days contributed by females was comparatively less. Of the persons normally outside labour force, the proportion of days contributed by females was higher than males.

TABLE 10. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED IN HARVESTING DURING THE BUSY WEEK WITH BREAKDOWNS FOR NORMAL OCCUPATION AND SEX

village	normal occupation												number of man-days
	agriculture			non-agriculture			outside labour force			total			
	male	female	total	male	female	total	male	female	total	male	female	total	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1. *G*	100.0	—	100.0	—	—	—	—	—	—	100.0	—	100.0	275
2. *N*	46.2	0.2	46.4	38.2	15.4	53.6	—	—	—	84.4	15.6	100.0	621
3. *C*	63.0	5.9	68.9	0.7	0.5	1.2	12.5	17.4	29.9	76.2	23.8	100.0	425
4. *D*	84.7	3.2	87.9	—	—	—	7.8	4.3	12.1	92.5	7.5	100.0	561
5. *S*	97.6	1.5	99.1	0.1	—	0.1	—	0.8	0.8	97.7	2.3	100.0	1201

4.5. *Source of labour supply.* The necessity of obtaining persons, other than household members, on hire or on exchange in the peak week was high-lighted in this study; the proportions of man-days worked by household members, interest holders (relatives and others who worked on some implicit understanding or for sympathy) and by persons obtained on hire and on exchange in the villages are shown in Table 11. In the "non-agricultural" villages *N* and the "near" villages *S*, the major share of the man-days was contributed by persons on hire (59 percent) and persons on exchange (56 percent) respectively; the contribution by household members ranged from 32 percent in the "near" village *S* to 76 percent in the "next remote" village *D*. The contribution by persons engaged on hire was also considerable in "remote-low population density" village *C* (47 percent) and in village *G* (32 percent).

TABLE 11. PERCENTAGE DISTRIBUTION OF MAN–DAYS WORKED IN HARVESTING DURING THE PEAK WEEK BY SOURCE OF LABOUR SUPPLY

village	source of labour supply					number of man-days
	household members	interest holders	on hire	on exchange	total	
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1. *G*	65.2	0.7	31.6	2.5	100.0	275
2. *N*	40.9	0.3	58.5	0.3	100.0	621
3. *C*	53.2	—	46.6	0.2	100.0	425
4. *D*	75.6	0.4	14.4	9.6	100.0	561
5. *S*	31.9	—	12.5	55.6	100.0	1201

4.6. The distribution of man-days in the peak week by local and outside labour in the five villages has been shown in Table 12. A major proportion of the persons obtained on hire and on exchange had normal residence outside the villages; the contribution towards the man-days in the peak week by outsiders was thus substantial in the villages where these two types of labour predominated.

TABLE 12. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED IN HARVESTING DURING THE PEAK WEEK BY TYPE OF RESIDENCE

village	local	outsider	total	number of man-days
(1)	(2)	(3)	(4)	(5)
1. *G*	94.9	5.1	100.0	275
2. *N*	51.2	48.8	100.0	621
3. *C*	78.8	21.2	100.0	425
4. *D*	91.1	8.9	100.0	561
5. *S*	77.6	22.4	100.0	1201

4.7. The distribution of man-days worked by the hired labour during the peak week by type of residence has been given in Table 13. The contribution by outside hired labour was maximum (83 percent) in "non-agricultural" village *N* and minimum (16 percent) in village *G*. The need of recruitment of outside hired labour at the peak period of harvesting is clearly manifest in the Table. Household members residing out[1] and interest holders contributed very small proportion to the total man-days worked by outsiders and they are excluded from consideration in this Table.

TABLE 13. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED IN HARVESTING DURING THE PEAK WEEK OBTAINED ON HIRE BY TYPE OF RESIDENCE

village	local	outsider	total	number of man-days
(1)	(2)	(3)	(4)	(5)
1. *G*	83.9	16.1	100.0	87
2. *N*	17.1	82.9	100.0	363
3. *C*	54.5	45.5	100.0	198
4. *D*	39.6	60.4	100.0	81
5. *S*	48.8	51.2	100.0	150

---

[1] The well-known phenomenon of the seasonal migration of factory labour was not observed in the Study, because perhaps of the population composition of the sample villages.

4.8. The contribution by migrants on hire during the peak week by type of migrant (from neighbouring or distant areas), normal occupation and sex has been given in Table 14. In the "remote-low population density" village *C* and in village *G*, all migrants on hire were obtained from neighbouring villages. In the "next remote" village *D*, migrants on hire from neighbouring villages constituted 86 percent of the man-days contributed by all migrants on hire. In "non agricultural" village *N* and the "near" village *S*, migrants on hire from distant areas contributed 88 and 99 percent respectively of the man-days worked by all migrants on hire. Only in "non-agricultural" village *N*, any contribution by migrants on hire with normal non-agricultural occupation was made; their contribution being 97 percent, 31 percent by females. Only in the "remote-low population density" village *C*, of the man-days contributed by migrants from neighbouring villages on hire, about 17 percent was made by persons normally out of labour force, 14 percent being made by females.

TABLE 14. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED IN HARVESTING BY MIGRANTS ON HIRE DURING THE PEAK WEEK BY TYPE OF MIGRANTS, NORMAL OCCUPATION AND SEX

type of migrant	normal occupation											
	agricultural			non-agricultural			out of labour force			total		
	male	female	total	male	female	total	male	female	total	male	female	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
				(1) village : *G*								
1. neighbouring village	100.0	—	100.0	—	—	—	—	—	—	100.0	—	100.0
				number of man-days—14								
				(2) village : *N*								
1. neighbouring village	1.7	0.3	2.0	5.0	5.3	10.3	—	—	—	6.7	5.6	12.3
2. distant village	1.0	—	1.0	60.8	25.9	86.7	—	—	—	61.8	25.9	87.7
3. total	2.7	0.3	3.0	65.8	31.2	97.0	—	—	—	68.5	31.5	100.0
				number of man-days—301								
				(3) village : *C*								
1. neighbouring village	71.2	12.2	83.4	—	—	—	2.2	14.4	16.6	73.4	26.6	100.0
				number of man-days—90								
				(4) village : *D*								
1. neighbouring village	85.7	—	85.7	—	—	—	—	—	—	85.7	—	85.7
2. distant village	8.2	6.1	14.3	—	—	—	—	—	—	8.2	6.1	14.3
3. total	93.9	6.1	100.0	—	—	—	—	—	—	93.9	6.1	100.0
				number of man-days—49								
				(5) village : *S*								
1. neighbouring village	1.3	—	1.3	—	—	—	—	—	—	1.3	—	1.3
2. distant village	98.7	—	98.7	—	—	—	—	—	—	98.7	—	98.7
3. total	100.0	—	100.0	—	—	—	—	—	—	100.0	—	100.0
				number of man-days—77								

4.9. *Intensity of work during the peak week.* For studying intensity of employment during the peak week, the man-days worked by persons directly engaged in harvesting were recorded in intensity classification of less than quarter, quarter, half, and more than half. The distribution of man-days in these intensity classifications by source of labour supply and the average intensity has been given in Table 15. The average intensity was very low (one-third) in the near village *S*; it varied from two-thirds in village *D* to three-fourths or more in the other villages, the maximum being in the "non-agricultural" village *N*. The diluted participation of household members and the high proportion of exchange labour in the "remote-low proportion of paddy area" village *C* were features to be noted.

TABLE 15. PERCENTAGE DISTRIBUTION OF MAN-DAYS WORKED IN HARVESTING DURING THE PEAK WEEK BY INTENSITY OF WORK FOR DIFFERENT SOURCES OF LABOUR SUPPLY

source of labour supply	percentage of days worked with intensity					number of man-days	average intensity
	less than quarter	quarter	half	more than half	total		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		1. village : *G*					
1. household members	—	2.8	10.1	87.1	100.0	179	0.77
2. on hire	—	4.8	8.4	86.8	100.0	83	0.76
3. all sources[1]	—	3.3	9.2	87.5	100.0	271	0.77
		2. village : *N*					
1. household members	—	—	7.6	92.4	100.0	249	0.79
2. on hire	2.1	—	5.0	92.9	100.0	341	0.78
3. all sources[1]	1.2	—	6.1	92.7	100.0	594	0.78
		3. village : *C*					
1. household members	—	9.0	11.2	79.8	100.0	223	0.73
2. on hire	—	6.2	7.2	86.6	100.0	195	0.76
3. all sources[1]	—	7.6	9.3	83.1	100.0	419	0.74
		4. village : *D*					
1. household members	—	2.4	53.3	44.3	100.0	413	0.63
2. on hire	—	—	30.8	69.2	100.0	78	0.72
3. on exchange	—	—	58.7	41.3	100.0	46	0.63
4. all sources[1]	—	1.8	50.3	47.9	100.0	539	0.65
		5. village : *S*					
1. household members	16.4	38.4	22.2	23.0	100.0	383	0.40
2. on hire	3.3	42.0	36.0	18.7	100.0	150	0.44
3. on exchange	19.2	54.1	20.6	6.1	100.0	640	0.30
4. all sources[1]	16.3	47.4	23.1	13.2	100.0	1173	0.35

[1] Including other sources of labour supply.

5. DIFFICULTIES IN CULTIVATION

5.1. *Difficulties experienced in cultivation during the aman season*—1954. Difficulties experienced during the current season were also recorded. Out of the 72 households eleven households usually under cultivation had to keep some land uncultivated in the particular season. In village *G* the reason for this was drought[1] (one household), in village *N*-flood (two households) and in village *S* flood[1] (seven households) and non-availability of bullocks (one household). On an average labour shortage of 19 and 13 man-days respectively was experienced by five households in village *S* and three households in village *C*.

6. DISCUSSION OF RESULTS

The necessity of obtaining labour on hire and on exchange during the peak week, local as well as from neighbouring and distant areas, was high-lighted in this Type-study as also the participation in harvesting by persons who normally have non-agricultural occupation or are normally out of labour force. It would thus be seen that if a percentage of population in the villages is withdrawn, there would obviously be some loss in the outturn of crops due to labour shortage in the peak week; especially when it is observed that eleven out of 72 sample households in the five villages experienced some labour shortage during the season and when it is known that harvesting cannot be extended beyond an optimum period without loss. The redeployment of the rural agricultural population and the consequential effects merit more intensive probes than the present material would permit. Conceptually, the whole problem hinges on the question whether the withdrawn population employed in another sector could contribute to the net national product in a better way. If it is granted that there would not be a proportionately greater loss of foodgrains by withdrawal of a part of the rural-agricultural population, two further questions pose themselves : (a) the loss of foodgrains would have to be compensated by imports so that the question of priorities would have also to be considered, which in its train brings in wider issues (Mahalanobis 1958, pp. 14–17); (b) this question apart, the provision of alternative, more productive work to the withdrawn population *pari passu* would not seem to be physically and concurrently achievable. If, on the other hand, it is assumed that there would be a net loss of foodgrains due to the withdrawal, the problem becomes more complicated. (The human and social costs of withdrawal have also to be considered in any such discussion of alternatives).[2] In any case, on balance of considerations, the only realistic solution is seen to be the injection of investments in the rural agricultural sector in the sphere of small scale and household production, which would mop up underemployment both seasonal and perennial, and the simultaneous broadening of the horizons of technological improvement for increased productivity, consistent with power planning: this, an essentially transitional phase in developmental planning, would also minimise the human and social costs of industrialization and is basically the procedure recommended by Mahalanobis (1955, pp.52-55).

---

[1] Severe drought hampered the sowing and transplantation of crops in the Southern districts and heavy rainfall and flood damaged large areas in the Northern districts of West Bengal during *Aman* 1954-55. *Supplement to the Calcutta Gazette*, April 21, 1955, p. 383.

[2] The inertia to be faced in any scheme of withdrawal is not generally taken into account : findings of the West Bengal Special Demography Study, conducted in 72 villages and 40 urban blocks in April-May 1954, and of the Try-out on Agricultural Labour Enquiry, 1956, covering 24 villages spread out all over India, suggests the existence of considerable inertia (Som and Bhattacharyya, 1958).

7. ACKNOWLEDGEMENT

The help of Ajit Das Gupta in this study is acknowledged and appreciated. We are also grateful to Asok Sen for valuable discussions.

REFERENCES

MAHALANOBIS, P. C. (1955) : The approach of operational research to planning in India. *Sankhyā*, **16**, 3–130.

——— (1958) : Science and national planning. President's Anniversary Address, National Institute of Sciences in India. Reprinted in *Sankhyā* **20**, 69–194.

SOM, R. K. and BHATTACHARYYA, G. C. (1958) : Some findings of the try-out on agricultural labour enquiry, 1956. *Studies Relating to Planning for National Development, PD. NSS/WP/4* (102), Indian Statistical Institute, Calcutta.

Paper received : January 1958.

NEXT STEPS IN PLANNING*

By P. C. MAHALANOBIS

Indian Statistical Institute, Calcutta

INTRODUCTION

India adopted planning for economic development with the initiation of the First Five Year Plan in 1951. At this stage, the plan was essentially a list of projects without any clear unity of purpose. A sense of strategy began to develop and certain broad aims were formulated in advance in 1954 for the preparation of a draft frame for the Second Five Year Plan. It was agreed that attempts must be made to get India out of the fear of unemployment in ten years or as soon as possible, to lay sound foundations for a continuing improvement in the level of living, that is of the national income; and also to bring about greater equality of opportunities.

1.2. In 1955 India accepted socialism as the goal of economic policy. The Second Five Year Plan took a long term view of economic development over a period of 10 or 15 years or more, and adopted a basic strategy for this purpose which involved a crucial change in the outlook. Three years have nearly passed after the initiation of the Second Plan in April 1956. There have been many gains and some set-backs; and much experience has been gathered. We must now prepare our mind for the Third Five Year Plan which will have to be started in about two years. It is, therefore, appropriate that we should now consider the next steps in planning.

REVIEW OF INDUSTRIAL DEVELOPMENTS

2.1 Before doing this it will be useful to consider previous developments in India in the context of the history of industrialization of the more advanced countries of the world. The industrial revolution started about two hundred years ago in the United Kingdom, the USA and countries of Western Europe where science and technology were highly developed. At first there was expansion of the manufacture of textiles and other consumer goods. The use of steel for industrial purposes started less than a hundred years ago. The world production[1] of steel was less than one million metric ton in 1870, and reached 4.4m. tons in 1880 and 28m. tons in 1900. In 1870 UK and Germany were producing about 200,000 tons per year and USA and Russia had just started. By 1880 UK and USA reached 1.3m. ton and Germany reached this level in 1885, and France in 1895. These countries were already industrialized and had a high level of living indicated by per capita income of from 400 to 600 US dollars per person per year (at current dollars for 1955 and 1956) at the time they reached the level of production of one million ton of steel per year. Health and education also were fairly well developed with the number of physicians ranging between 3.6 and 16.3 per ten thousand of population. Developments were fairly rapid from this time and USA reached 4.3m. tons in 1890, Germany 4m. tons in 1895, UK 5m. tons in 1900 and France 4.7m. tons in 1913. During this period there was some further expansion of health, education and research which was broadly in accordance with the previous trend.[1] An output of 4 or 5 million tons of steel per year usually makes a turning point in industrial progress. At this level it becomes possible to establish comprehensive manufacturing complexes of steel, heavy machine building and heavy electrical industries, and chemical engineering.

---

* Anniversary Address at the National Institute of Sciences of India by the President, Professor P. C. Mahalanobis on 20 January 1959.

[1] Data on production of steel are given in Table 3 in the Appendix.

2.2. Japan had started producing steel just before the First World War and reached one million ton in 1923 when Japan's per capita national income was still quite low and only 125 US dollars. Health and education in Japan were, however, as highly developed as in the advanced countries of Europe and America. Japan reached 4.9m. tons in 1935, that is, in 12 years. It is important to note that modernization in Japan had started about 90 years ago with the Meiji Restoration with deliberate emphasis on health, education, scientific research and industrialization. Japan, for example, had made primary education free and compulsory before UK.

2.3. Russia had reached the production of about one million ton of steel as early as 1895 but was quite backward in education and health with 1.6 physicians per ten thousand of population and with only 3 percent of the population in primary schools. Owing to war, revolution and civil war production in USSR had decreased to 0.6m. ton in 1923, but recovered to 4.9m. tons in 1929 by which time education and health had also developed in a remarkable manner approaching conditions prevailing in the more advanced countries at corresponding epoch. China started very recently with 0.6m. ton of steel in 1950 and rapidly reached 5m. tons in 1956 and 10m. tons in 1958

2.4. Until recently industrial developments in India were taking place rather slowly and in a somewhat imitative way of the earlier history of industrial growth in Europe and America in which we can distinguish four layers or phases. On the top or the surface comes consumer goods which can be established fairly quickly with the help of imported machinery. The second layer is the production of electricity, steel and other metals; this is highly capital intensive and requires large imports of machinery and also an appreciable amount of time. Next is the manufacture of heavy machines, heavy electrical apparatus and chemical engineering (for the production of industrial raw materials). This usually requires a much longer time but, once established, would enable a large part of the capital and producer goods being manufactured out of domestic resources. Even more basic is the expansion of health, education, science, and of scientific and technical manpower which takes practically a generation to mature.

2.5. Following the above pattern only upto the first phase of consumer goods, India had become the second biggest producer of textiles in the world and had also developed the manufacture of other consumer goods without however establishing the basic industries like metals, heavy machine building and heavy electricals. In fact, in India the production of steel had remained basically at a little above the level of one million ton after nearly half a century of the beginning of the industry, although the country has very large resources of iron ore of the best quality. The economy was developing only very slowly on the basis of immediate demands and supplies without any long range view. Also, there was comparatively little development of health, education, and science in relation to the size of the country and its population.

---

[1] Information on certain characteristics of development is given for selected countries in Table 1 in the Appendix.

GAINS IN THE SECOND FIVE YEAR PLAN

3.1. There were important gains in the Second Five Year Plan which marked a radical break with the old way of thinking. It was realised that unemployment and poverty can be cured only through rapid industrialization. It was not enough to expand the production of textiles and consumer goods. It was necessary to increase the supply of energy and of machinery and tools of production to create more employment. It was realised that the only way to do this in a big country like India was to establish, as rapidly as possible, the basic steel, heavy machine building, and heavy electrical industries and chemical engineering. This would make it possible to produce capital and producer goods of all kinds and thus expand industrial investments increasingly out of domestic resources. This is the only way to secure increasing supplies of machinery and energy, with which we can create more employment and also can give the workers more machinery so that the total production in the country can rise rapidly.

3.2. In India the highest priority must be given to the establishment and expansion of the basic industries. Large investments in the basic industries would, of course, give rise to an increasing demand for food, cloth and other consumer goods. In India it is possible and desirable to meet this demand by setting idle hands to work in the traditional way or in small scale industries to produce consumer goods and other necessities as much as possible. This would also create much employment all over the country.

3.3. It was appreciated that, in India the production of an agricultural surplus is the key to industrialization. It is not only essential to grow enough food and fibres for our own requirements but it is also necessary to produce a surplus in the form of either industrial or food crops. In India agriculture and manufacturing industries are completely interlocked. Economic progress depends on the advance of both. Advance of one step in agriculture would supply food and raw materials for advance of one step in manufacturing industries which again, in its turn, would speed up irrigation and increase the supply of fertilizers and pesticides and help in the promotion of scientific research, which would lead to further advances in agriculture. The importance of the scattered or diffused phase of production, both in agriculture and in traditional and small scale industries, was broadly appreciated. It is also being increasingly realised that land reform and the organization of village co-operatives are urgent and essential needs for increasing agricultural production.

3.4. The Second Five Year Plan accepted the above strategy of economic development in a general way. It was decided to expand steel production to about 5 million tons; to give high priority to atomic energy; to establish a heavy machine building industry and the manufacture of heavy electricals; and also to develop prospecting for oil exploration and to establish oil refineries of our own. India would be now able to start developing large manufacturing complexes.

3.5. There were, however, gaps and set-backs also. An acute shortage of foreign currency occurred in the very first year of the Second Plan which in my opinion was not inescapable but was largely due to lack of adequate tools and instruments of planning and also partly to lack of proper appreciation of the strategy of the Second Plan. The shortage of foreign currency created great difficulties. I discussed this question in my Anniversary Address last year. The remedy is in our own hands because in India we can manufacture capital and producer goods increasingly out of our own domestic resources, once the basic industries are established and developed. Therefore the only radical cure of the shortage of foreign currency would be by building up the basic industries as rapidly as possible.

3.6. Foreign aid, especially in the form of long term loans for capital goods and technical personnel, would be of critical importance at the stage of what P.M.S. Blackett has called "assisted take of" in industrialization. Indefinite reliance on foreign loans is neither possible nor desirable. A continuing stream of short-term financial loans from abroad would not be of much help in the long run and may make the position worse. The chief source of capital formation must be domestic savings. Fortunately, the need of raising more resources is admitted and gradually action is being taken in this regard. One notable advance was the acceptance of the principle of an expenditure tax together with a tax on wealth.

3.7. In the highly developed countries of the West, taxes on commodities are usually looked upon as "regressive" as being a burden on the poor. Public enterprizes are also expected to be run on a no-loss no-profit basis. Fortunately our outlook is changing and it is being realised that in an underdeveloped country like India excise and customs duties, purchase tax on commodities or a levy on services would be convenient and adaptable methods to raise resources. It is also agreed in principle that public enterprizes should earn and contribute increasing returns for purposes of national development.

3.8. It is encouraging that the question of scientific and technical manpower is receiving increasing attention. The subject is being studied at a technical level in the Division of Perspective Planning in the Planning Commission. A Standing Committee of the Cabinet has been set up for manpower; and administrative units have been established in the Central Government and in the States.

Weaknesses in planning

4.1. The Second Plan, no doubt, has been a great step in advance. It has set production forces in motion which cannot possibly be reversed without disaster. There are, however, serious gaps and weaknesses.

4.2. Review of industrialization has shown that there is a close connexion between economic development and progress in health and education. (Some relevant data for both developed and underdeveloped countries have been given in Table 2 in the Appendix). In West Europe and America there was simultaneous progress of industries and health, education and science. In Russia, before the first World War, the production of steel had reached 5 million tons but there was little progress in health and education; and there was revolution. In USSR, China and other socialized countries and also in Japan great emphasis was and is being given to progress in health, education and science *pari passu* with industrialization. India would soon reach the critical level of steel production of 5 million tons but is still backward in health, education and science.

4.3. Before considering these three points in some detail I shall refer to another disturbing aspect of the present situation, the steady increase in unemployment.

INCREASE IN UNEMPLOYMENT

5.1. We do not yet have in India any regular reporting of unemployment. Some indications are however, available from trends in the number of applicants in the Labour Exchange Registers in India for which some data are given in Tables 4 and 5 in the Appendix. It is disturbing to note that the total number of persons registering in these Exchanges had more than doubled in five years between January 1953 and December 1957. It is also worth noting that the Labour Exchanges can find employment for only a very few; and the average number of applicants for each notified vacancy increased steadily from 23.3 in 1953, year by year, to 33 in 1957. Also as all notified vacancies were not filled by applicants recommended by the Labour Exchanges, the number of applicants for each effective placement was nearly 40 in 1947 and had steadily increased to nearly 51 in 1957. That is, the Labour Exchanges can find employment for only two persons or less out of every *hundred* applicants, on an average. This is not a criticism of Labour Exchanges but simply an indication of the heavy back-log of unemployment.

5.2. This back-log differs in different occupational categories for which some information is given in Table 4. The ratio of applicants to effective placements increaesed in practically all categories except in "others." The greatest relative increase occurred among those seeking "education," "clerical" and "unskilled" jobs. In 1957 there was only one effective placement for 36 applicants seeking educational posts; and one for 75 seeking clerical or unskilled jobs. It is clear that the size of the second plan was not big enough to absorb the fresh seekers of employment who joined the labour force through the growth of population. A much bigger plan is essential to get rid of the fear of unemployment.

5.3. In my Anniversary Address last year I gave some information on unemployment and underemployment obtained through the National Sample Survey. I explained the difficulties of measuring underemployment and stated that there was scarcely any doubt that 10 or 12 million persons in India were unemployed or severely underemployed and that this figure might rise to even 25 or 30 million if consideration is extended to those who were sitting idle for more than half of their normal working time. I had drawn attention to the fact that about 10 percent of educated persons (matriculates and above) were probably unemployed, and that this number was likely to rise. The evidence from Labour Exchanges would seem broadly to corroborate this.

5.4. It is clear that the Second Five Year Plan has not succeeded in absorbing, in sufficiently large numbers, the new seekers of employment. It is also likely that unemployment is continuing to increase among educated persons especially in urban areas. This is a disturbing sign. The only conclussion to be drawn is that the Third Five Year Plan must be much bigger in size than the Second Five Year plan. It would be remembered that in the Second Five Year Plan it was visualized that net investments should reach Rs. 9,900 crores in the Third Plan. Let us round off the figure, and adopt Rs. 10,000 crores as the target for investment in the Third Five Year Plan.

HEALTH

6.1. Health, education and research have indeed a dual role. These are no doubt significant constituents of the level of living and, in this sense, are fruits of national development. On the ohter hand, advance of health, education and research are of basic importance in bringing about industrial and social progress. I shall first consider health.

6.2. In India at present there are about 70,000 physicians of whom a little less than a half are trained at the University level and about 37,000 are licenciates with four-year training in junior medical schools. The distribution of doctors by level and qualification and state of residence in 1954 is given in Table 6. The number of doctors in India as a whole is about 176 per million of population or one physician for about 5700 persons on an average. The number however varies widely from one state to another, from over 600 per million (or one for 25,000) in Rajasthan and Manipur. What is, however, much more serious is that most of the doctors reside in urban areas; and the total number of doctors residing in rural areas (with a population of about 330 million) may be below 10,000 that is, only one doctor for 30,000 persons or more on an average. In some rural areas it is known that there is only one qualified doctor for 50,000 or 100,000 persons.

6.3. An important committee on health and medical care (popularly known as the Bhore Committee from the name of its Chairman) stated in the foreword to its report submitted in 1946 that "*no individual should fail to secure adequate medical care because of inability to pay for it.*" This Committee had prepared a long-term programme most of which still remains unimplemented.

6.4. There is no difference of opinion regarding the desirability of providing medical care to all who need it. In India, unfortunately, in some respects we are still at the stage of a somewhat superficial imitation of the most advanced countries of Europe and America. Junior medical schools were abolished after independence on the view that our countrymen cannot be allowed to have anything less than the best. And yet to provide a sufficient number of university trained doctors and adequate hospital and medical services upto the standard of the rich countries of the world is absolutely impossible at present.

6.5. Doctors and drugs are both extremely scarce in India. Doctor's fees and the price of drugs are also high which keep these beyond the reach of most people. The extreme inequality in the distribution of expenditure on medicine and medical services can be appreciated from the relative share of expenditure on these items, shown below (and also from the Chart in the Appendix) for the bottom ten percent, the lower half, and the top ten percent of all households.

percent of households	cumulative percentage of expenditure on			
	medical services		medicine	
	USA[1] 1950	India[3] 1955	India[2] 1952	India[3] 1955
(1)	(2)	(3)	(4)	(5)
	percent share of expenditure*			
bottom 10 per cent	5	0.1	1	1
(lower half) 50 ,,	37	6	9	8
top 10 ,,	19	64	45	56

[1] *USA* 1950 : Studies on Consumer Expenditure, 1950, Vol. VIII (University of Pennsylvania).

[2] *India* 1952 : National Sample Survey : 4th Round; April-September 1952: All-India (Urban).

[3] *India* 1955 : National Sample Survey : 9th Round; May-November 1955 : All-India (Rural).

**Note :* The classification is by size of income of households for the USA 1950 data; by size of expenditure of households for India 1952 data; and by size of expenditure of persons for India 1955 data. The American and Indian results are therefore not comparable in a completely rigorous way but the broad picture is clear. The expenditure on both medicine and medical services are far more concentrated in India showing that it is only the very few rich who can afford to pay for medicine and medical services. The two series for India 1955 have been, however, classified in the same way; and show that the distribution of expenditure on medical services is relatively more concentrated, that is, medical services are even more scarce than medicine in rural areas of India.

6.6. The above Table gives some similar figures for urban areas in USA in 1950. Looking at the first line of the table it is seen that the bottom ten percent of American households share five percent of the total expenditure on medical services incurred by all households; the corresponding share of the bottom ten percent of rural households in India is negligibly small and only about 0.1 percent. From the second line it is noticed that the lower half of American households share 37 percent of the total expenditure while the Indian households share only 6 percent of the total. We may also look at the top. Ten percent of the households at the top in America take a share of 19 percent of the total while the corresponding Indian households at the top have 64 percent or nearly two-thirds of the total expenditure on medical services. The position is almost as bad in the case of medcine. In India the lower half of rural households share only 8 percent of the total expenditure on drugs,[1] the corresponding figure for urban households being 9 percent.

6.7. It may be pointed out, at this stage, that in countries with free and universal health services like UK and USSR every person, in principle, receives essential medical services and medicine. In both countries, as private medical practice is permitted and drugs can be purchased in the market, some medical expenditure may be incurred by the rich on their own, but there is no discrimination between the rich and the poor so far as the medical care is available to the poor either free of charge or at a very small cost.

6.8. I should like to make a few suggestions for consideration for the Third Five Year Plan. Firstly, I think it would be wise to reintroduce training extending over, say, 3 years for a junior type of doctor or auxiliary health workers who would be prepared to serve in villages for a modest allowance of say something like Rs. 1,000 per year with permission for private practice within certain limits. The training cost would be possibly one-fourth or one-fifth of that of a fully qualified doctor. The pay or allowance would also be one-fifth or even less of the pay which would have to be given to a university level doctor. The cost of training and maintaining a junior doctor may be thus five times cheaper.

6.9. The organisation of medical teaching at a junior level would also be feasible on a much larger scale than at the university level. There are about 400 district hospitals, which, with some additional facilities, can serve as training centres for junior doctors; and which, on an average, may be able to turn out 25 trained persons each per year which would supply 10,000 junior doctors per year after the scheme is in full operation. This would make it possible to have, one junior doctor for a group of ten villages in roughly half a million villages in eight or ten years. One fully qualified doctor may be placed in charge of about five junior doctors; and should be able to give some attention to the more difficult cases.

6.10. The cost of providing hospital and other health services would also have to be brought to a level which the country can afford. The cost of construction standard and specifications which would be appropriate for a poor country like India; and, secondly, by improving the efficiency of public construction. It is also necessary to develop rapidly the manufacture of drugs and medical goods, and also to establish factories for the manufacture of machinery for this purpose. There, is of course, need of more systematic research for a fuller exploitation of our own natural resources for medical and health purposes.

---

[1] Similar figures for the distribution of consumer expenditure on other items like foodgrains; all food items, cloth, sugar, education, and medical services are given in Tables 12, 13 and 14 in the Appendix.

6.11. Finally, it seems to me it would be very wise to initiate a truly national health service, even if this be in a skeletal form, during the third plan period. The emphasis in the first instance should be on establishing a large number of small village units which would provide essential medical and health service to the villages and also serve as family planning centres. This is the only way in which there would be any real chance of promoting birth-control effectively. This is a matter of great importance and urgency.

EDUCATION

7.1. India is backward in education generally. Some comparative information is given in Table 7 for selected countries. In most advanced countries there is practically no illiteracy; and in China also the position is improving rapidly. India is still far off from universal and compulsory primary education. In India the number of illiterates (including both males and females) was 83.5 percent of the population according to the Census of 1951. The National Sample Survey indicated that there was only a slight reduction of illiteracy to 82.5 percent of the population as a whole in 1955. The number of students in primary schools in India was about 60 per 1000 of population in 1956 against 87 in China in 1955, 144 in Japan in 1956, 125 in USSR, and 154 in USA.

7.2. The position is quite different at the secondary level. India had 23 students per 1000 of population at the secondary level in 1956 which was three times higher than the corresponding figure 7.4 in China in 1955. Also, the Indian figure does not compare too badly with 38 in USSR in 1957-58 or 48 in USA in 1955. The number of students in the university and equivalent levels in India is remarkably high with a rate of 1.8 per 1000 of population which can compare favourably with 0.5 in China, 1.6 in UK, or 2.5 in West Germany and 3.5 in France.

7.3. This structural difference in education comes out even more clearly if we adopt the number of primary students as the base for comparison. For every thousand students in the primary stage, India had about 384 at the secondary level in 1956 which was much higher than the corresponding proportion of 84 in China, and also higher than the proportions in France (264), USSR (300), and USA (312). The number of students in the university and equivalent levels per thousand students in the primary level is also very high in India and was about 30 against 5 or 6 in China, 15 in UK, 26 in West Germany, 31 in France and 47 in Japan. Only USA and USSR have higher proportions.

7.4. The above comparisons bring out several points. India is weakest at the primary level and is far behind China and all advanced countries of the world. India has a much stronger position at the secondary level on a population basis. If however, the number of primary students is used as the base for comparison then India has proportionately larger numbers in the secondary schools than most advanced countries of the world including USA and USSR. At the university and equivalent levels, also, India has a higher proportion of students than UK and has a proportion of the same order as Germany or France on the basis broadly of both population and of the number of students at the primary level.

7.5. It is important to note that higher education in India has been expanding quite rapidly in recent years. Some relevant information is given in Table 8 which shows the enrolment of students by level of education, subjects, and sex in 1950-51 and 1956-57. It can be seen from this table that the total enrolment at higer levels (from intermediate to post-graduate and equivalent courses) was about 400,000 in 1951 and increased to about 780,000 in 1956-57, that is, the number had almost doubled in six years. It can be further noted that the largest increase had taken place at the intermediate level. On the whole, there was relatively greater expansion in arts, commerce and law compared to science and technology including medicine and veterinaries.

7.6. The outturn of qualified persons in India, at intermediate and higher levels also showed a large increase and, more or less, kept pace with the increase in enrolment. Relevant data for the period 1951-1956 are given in Table 9 in the Appendix. There was, however, a good deal of wastage. On the whole, only about half the candidates who appeared at the examinations at the higher (university and equivalent) levels succeeded in passing these examinations. The proportion of successful candidates was the lowest at the intermediate stage (between 42 and 46 percent). Relevant figures on number and percentage of successful candidates are given in Table 10. Large percentages of failure give rise to much frustration and to much waste of time, money and effort.

7.7. Some interesting information can be obtained by examining the number of students in higher courses in proportion to the number of matriculates for which relevant statistics are given in Table 11. It can be noticed from this table that for every 10,000 students who matriculated in 1952, about 6,750 had enrolled for courses at the intermediate level in 1952-53 out of whom about 4,400 succeeded in passing the appropriate examinations at this level in 1954. The enrolment at the degree levels was 3,319 in 1954-55, and the number receiving degrees two years later in 1956 was 2,364. The enrolment in post-graduate courses in 1956-67 was 1,552. It may be noted that for every 10,000 candidates who had matriculated in 1952 only 2,920 or less than 30 percent took up courses in science and technical subjects at the intermediate level; and less than 12 percent proceeded with courses in science, technology, agriculture, medicine and veterinary at the degree stage. Finally, only 156 or less than two percent took post-graduate courses in science and technology. The educational base is poor for a rapid expansion of scientific and technical personnel at higher levels.

7.8. Higher education in India developed to a great extent in a somewhat imitative way on the British pattern. Social and economic conditions are however entirely different in India. Academic ideals which are realisable in the UK sometimes necessarily degenerate into superficial imitation and window dressing. Indian colleges often have very large numbers of students who are not adequately prepared for higher education. University and college teachers are poorly paid and are often obliged to do a good deal of part-time outside work. Classes are very big, and library and laboratory facilities are often poor. It is not surprising that conditions of work in the universities are not entirely satisfactory.

7.9. In the advanced countries of the world there are usually good opportunities for productive employment for the students who come out of universities and other higher institutions. In India the position is entirely different. Only a very small proportion receive vocational or professional training. A very large number receive what is called general education which does not make them fit to undertake any particular productive work. As I have already mentioned, unemployment among educated persons in India is increasing.

This is creating a great deal of frustration and discontent. The greatest weakness in India is the lack of an organic relation between the system of education and national needs for economic development.

7.10. A radical reform of the educational system is a most urgent next step in planning. It is necessary, to turn out an increasing number of suitably trained personnel to meet the increasing demands of an expanding economy. The first thing in education must therefore be to give appropriate vocational and professional training which would make graduates of higher educational institutions qualified to undertake productive work of one kind or another. This is not in any way incompatible with what is called general culture. In fact, the two can very well go together. Dr. Zakir Hussain in his Patel Memorial Lectures has recently given a most illuminating exposition of this theme.

PLANNING FOR SOCIALISM

8.1. At this stage it is desirable to recall that in India socialism has been declared to be the goal of economic policy. It is quite proper not to try to give any rigid definition. And yet it is essential to have some clear understanding in this matter as a guide to action. In USSR, China and other countries in which socialism of one form or another has been already established, planning is considered to be an essential requirement for economic and social progress; and the aims and objectives of planning are settled in accordance with the principles of socialism in the accepted form. In other words planning in the socialized countries is an instrument of socialism. In India the position is different. Socialism has been accepted as the aim but has not yet been established. It is, therefore, important to examine the relation between planning and socialism.

8.2. Many different ideas have been associated with the concept of socialism in different countries and at different times. I shall select three ideas which have special relevance to our problems. The first is rapid economic development through full utilization of the resources of the country. This was the object for which planning was established in India and should continue to receive full attention. The second idea I should like to mention is removing progressively large disparities in wealth, income, and power; and the third is fostering equality of opportunities of all kinds for all the people of India.

8.3. The distribution of wealth, income and privileges is extremely unequal in India. The number of persons who pay income-tax is of the order of five lakhs or half a million.[1] Multiplying by five, (which is the average number of persons per household) the total number of persons in families paying income-tax would be about 2.5 million out of a population of nearly 400 million or appreciably less than one percent. The class which has some money and influence in India is thus extremely small and possibly forms not more than one or two percent of the whole population. Facilities for education being both scarce and expensive, higher education is probably almost a monopoly of this privileged class at the top. The inequality in the distribution of expenditure on educational services can be seen from the accompanying Chart and Table 14 in the Appendix. The bottom 10 percent of rural households in India have a share of less than one percent of the total educational expenditure and the lower half (50 percent) of households shared 11 percent only; while the top 10 percent have as much as 36 percent. Many poor people do not get even the opportunity to learn how to read and write. Many who complete the primary stage cannot join secondary schools;

---

[1] The actual number depends on the level of income at which income-tax becomes chargeable ; this level has fluctuated between Rs. 2500 and Rs. 4200 per year in recent years.

many who complete the secondary stage are unable to go to institutions of higher education on account of poverty. The same process of selection, on the basis of the income of the parents, continues at all levels of higher education. Finally it is only the rich people who can send their children abroad for higher training and education.

8.4. Inequalities in medical care result in a larger number of deaths among the poor people. Inequalities in educational opportunities have a deeper and more pervasive effect. It is true that some extremely able students succeed in securing scholarships upto the highest level but their number is very small. By and large, it is rich people who have the opportunity of giving their children the type of education required for posts of influence and responsibility in the country. By and large, those having such training are selected for posts of responsibility on the strength of their higher educational qualifications. In this way the power and privileges of a small group of people at the top tend to be not only preserved but strengthened. In both public services and organized private enterprizes practically all posts of influence and power are held by persons belonging to the same small privileged class. This has created an influential group of people who naturally desire to maintain their privileged position and power. During the British period many of the influential people were not enthusiastic about the political change because they were afraid of losing their own privileges. In the same way it is not surprizing that there are people in India today who are not enthusiastic about a rapid economic progress out of a similar fear of losing their privileges and power. It is necessary to remove barriers to educational opportunities to overcome such difficulties. This is the only way in which a sound foundation can be laid for democracy and socialism in India.

EDUCATIONAL REFORM

9.1. The only real remedy is to make education entirely free and also to ensure that no deserving student would be deprived of educational opportunities on account of poverty. It is no doubt necessary and desirable to do this at all levels of educations; but it may be too difficult a task to undertake immediately. As a first step, it is suggested, that this reform should be carried through at the university and equivalent levels during the period of the Third Five Year Plan. This would call for several types of action. Firstly, all fees at the university level would have to be abolished by increasing the direct contribution from Government by an equal amount. Secondly, enough financial assistance would have to be provided by Government to enable every deserving student being maintained during the whole period of his study in higher educational institutions. Thirdly, selection for admission to higher education would have to be made entirely on merit. Fourthly, it would be desirable to institute evening classes, correspondence courses, and external examinations on a large scale to provide opportunities for higher education to those who are unable to attend the day-classes and also to maintain a policy of open door as a safeguard against rigidities and defects which may develop within a strictly controlled system of admissions. Finally, it would be desirable to initiate a scheme for constructive work for the nation which would be obligatory for one or two years for all students who are admitted to higher educational institutions or who graduate from such institutions. This is not the occasion to give any detailed scheme but it may be useful briefly to consider some of the implications of the above proposals.

9.2. We have seen that total enrolment in higher education had nearly doubled in six years and was about 780,000 in 1956-57. The direct expenditure on higher education was about Rs. 293 million (Rs. 29.3 crores), out of which about 42 percent or Rs. 120 million

(Rs. 12 crores) came from students' fees.[1] The number of students during the Third Plan period would no doubt be higher. If expansion continued at about the same rate, the number in the middle of the Third Plan period may reach a million and a half. It may not be however necessary or desirable to allow such an increase in future, and we may provisionally accept a target of one million students in higher education in the Third Plan but also keep in mind a million and a half. For one million student, the Government grant in lieu of fees would come to about Rs. 16 crores, and for a million and a half to about Rs. 44 crores per year at current rates and prices.

9.3. As regards financial assistance for maintenance, at the rate of Rs. 300 or Rs. 400 per year per student on an average, the cost would come to Rs. 30 or Rs. 40 crores per year for one million students. I have used average figures because it would be quite reasonable to adopt something like a "means test" so that the amount of the stipend would depend on what assistance the parent would need to maintain the student without any strain. Secondly, stipends of a larger amount may be offered as incentive to attract students to science and technology or other suitable subjects and also to post-graduate course. The total cost may be of the order of Rs. 50 or Rs. 60 crores per year. If the Third Plan succeeds in raising the national income by five percent per year (and this is about the lowest safe limit) then the additional income generated would be at least of the order of Rs. 600 crores per year or more out of which there should not be any difficulty in providing the above expenditure.[2]

9.4. Selection for admissions on merit is indeed a formidable task, but I venture to think it is not beyond our capacity if we call science to our aid. The objects of the admission test must be clearly defined, and attempts must be made to maintain uniform standards throughout the country. I shall make a few suggestions to serve as a basis for further discussions. It may be desirable to try to assess such aspects of the candidate's qualification, for example, as (a) his content of knowledge in appropriate subjects, (b) his ability to search and compile relevant information from books and written materials, (c) his capacity to organize his knowledge and convey his ideas in a clear way in a written form, or (d) his skill in manual or technical work or ability to make scientific observations or experiments. These are all amenable to assessment on a more or less objective basis. For example, modern objective tests, in which a candidate has only to indicate his choice between different alternative answers, can be used in testing knowledge in a standardised manner throughout the country. The ability to compile and collect information from given books and materials can also be standardised at different levels of difficulty. The ability to convey in a coherent form his knowledge and ideas can be tested by asking the candidate to prepare written notes on the basis of material given to him or to be collected by him from given books and papers. His power of expression can be tested by the usual type of free essay for which several hours can be allotted. Some kind of practical work or experiments can also be made a part of the examination.

---

[1] Compared to a fee income of 42 percent of the total expenditure in India the income from fees at the university level is less than 11 percent in UK and is about 8 percent in the land-grant institutions many of which are the great State universities of USA.

[2] It is appropriate at this stage to point out that the proposals to abolish all fees in higher education and to award maintenance stipends out of government funds would merely involve a redistribution of the national income; and should not, in principle, cause any additional strain on the economy as a whole. It would be necessary, of course, for Government to raise the required amount through additional taxation or in other ways. But, in principle, the amount which was being spent by the parents of the students would be set free and would be available for either savings or consumption. In other words, there would be no additional call on physical resources.

9.5. I should like to mention that, in my opinion, modern statistical methods can be used very effectively to select a required number of questions for any particular paper out of a given "universe of questions" in a random manner and at the same time to maintain assigned levels of difficulty and subject coverage within prescribed limits with objective certainty. This would have two advantages. It would enable uniform standards being maintained throughout the country and from year to year. I would also permit almost infinite variations in the content of individual question papers so that copying from an answer book can be made practically impossible. Furthermore, with the help of modern electronic computers it would be possible to mark the objective type of papers very quickly and accurately.

9.6. I may also explain that the test for admission need not be one single examination but may very well be different examinations for different groups of subjects. For each examination it would be, however, necessary to maintain as uniform standards as possible over the whole country. To maintain parity or equivalence of standards between different examinations would be technically a more difficult task. However in principle, it is not absolutely necessary that the quality of the candidates should be the same in all subjects. If there is greater competition in certain subjects it is likely that such subjects would attract candidates of a higher quality on an average. This need not be prevented.

9.7. One great advantage of the proposed system would be that it would be possible to regulate admissions to particular subjects in accordance with social needs and opportunities of employment which are likely to be available to the students after graduation. In this way admission to higher education would be geared to future requirements of trained manpower.

9.8. It would be of advantage in many ways to institute evening classes, correspondence courses and external examinations to offer fuller opportunities for higher education. Correspondence courses are being used very extensively in USSR apparently with great success; and it would be possible to take advantage of Soviet experience. Arrangements may have to be made for laboratory or other institutional work for prescribed periods or during vacations. All this would, of course, call for much organizational effort and experimentation which would offer scope for much employment of persons with higher education. A standard system of external examinations (including practical work or tests as necessary) would keep the door open for every one and would be a safeguard against abuses developing within a closed system.

NATIONAL SERVICE

10.1. If the State undertakes to make education available at its own expense to all deserving students then it would be entirely proper to ask the students to undertake productive and constructive work for the country for one or two years. The maintenance stipend would be continued or can be increased in suitable cases during this period. In many countries of the world a period of service in the armed forces is compulsory. In India, instead of conscripted service in the defence forces, we may think of a period of obligatory service in peace camps or in a peace force for constructive work.

10.2. It would be probably convenient to ask the students to put in their prescribed period of service immediately after their graduation. This would supply every year a large number of highly trained persons who would be able to undertake productive work of various kinds. This would be of direct gain to the nation. For example, it would be possible to use a large number of such students to do teaching work at different levels all over the country. The medical graduates can help in bringing medical care to the villages. For the trainee himself it would be a valuable period of apprenticeship training, something like the "internee" work in hospitals by medical graduates, and would be of help in their finding employment suited to their training and aptitudes. The most valuable aspect of the peace camps or peace forces, if I may use this phrase, would be the bringing together of young persons from widely differing types of families in a comradeship of active work for the country. This would be an effective way of promoting the spirit of democracy.

10.3. Although I have spoken of students joining the peace camps or the peace force after graduation from higher institutions of education it is also possible to have "world schools", as Dr. Zakir Hussain has called them, for students before they come to the University. In USSR a new policy has been adopted of asking all students to work for at least two years from the age of 15 before they are admitted to institutions of higher education. Various possibilities are thus open. The really important point is to accept the principle that a prescribed period of service in a peace force for say two years at some stage or other would be an obligatory requirement for all educated persons.

Science and research

11.1. The importance of science and technology for national development has been recognised in India, in principle, since independence. It cannot be said, however, that progress has been entirely adequate or satisfactory. There is urgent need of improving the teaching of science at all levels. Much improvement is also needed in the organization of research. There is no time to make a comprehensive review, and I shall make some brief observations on some selected aspects of the problem.

11.2. Science is developing so fast that it is becoming increasingly difficult for teaching to keep pace with the advancement of knowledge. J. D. Bernal has recently pointed out that :

> "With a rapidly growing civilization, the young will year by year have more to know and also, will be required, by the new demands made on them, to know more... Something must be done to simplify and reduce the amount of learning at every stage... In the past with some reason, teaching was considered to be the passing on of known and established truths from one generation to the next... But clearly, in a civilization where the whole basis of knowledge is expected to change several times in a generation, the passing on of established knowledge becomes palpably inadequate and, also, from its over increasing bulk, impossible. It is lucky if most of what is taught is true, or supposed to be true, at the time it is taught. It is certain that it will no longer be considered true by the time those who are taught are half-way through their lives. What is needed, therefore is a different and lightened content of education. The emphasis will be on discovery rather than knowledge. It will be, not so much the passing on the established truths, as showing the way to criticize and discover new truths; in other words, the active part of the scientific method."[1]

---

[1] J. D. Bernal : *World Without War*, 1958, pp. 197–198.

11.3. Some serious thinking has been going on for some time in the Indian Statistical Institute on organizing an integrated course of science teaching in which emphasis would be given to develop the student's skill in observation, counting, measurement, logical analysis and interpretation of the facts and the design of new experiments to test the working hypothesis. One single set of observations and experiments may cover different branches of science, which, it is hoped, may give an integrated outlook on scientific method. It is no longer possible to have encyclopaedic coverage of knowledge. The only possibility is to try to make an intensive and integrated study which would touch particular aspects of different branches of science. I have no time to elaborate these thoughts but it seems to me that it will be well worth making some experiments in this direction.

11.4. It is essential to spread the teaching of science extensively at the school level without which it will be never possible to secure a sound foundation for advancement of science in India. Sir Alexander Fleck, in his presidential address to the British Association for the Advancement of Science in 1958 mentioned that in USSR children from 11 to 17 years devote 46 percent of their school time to science subjects. I am aware that it is not possible to approach this level immediately in India but we should keep some such target in view. Teaching of science, both in schools and universities, deserves serious attention by senior scientists.

11.5. As regards the organization of science in its wider sense, I have on other occasions pointed out that our own National Institute of Sciences is not very active. Our Institute is supposed to have responsibility of advising Government on scientific matters. In practice there is very little activity of this kind. This I think is partly due to lack of clear ideas regarding what should be the proper function of scientific societies in the changing conditions of the present time; and also, I am sorry to say, partly due to lack of a sense of responsibility on our own part. I have made some suggestions in my Annual Review in 1958 and need not say anything more on the present occasion.

11.6. I am a strong advocate of the public sector but this, in my view, does not mean concentrating all activities in government departments. On the contrary, in the scientific field, I think it is essential to build up and maintain truly independent scientific organizations which would be in a position to offer advice and constructive criticisms and suggestions to Government. I have already mentioned that we ourselves have not been sufficiently serious regarding our own responsibilities. But I also think that the time has come for Government to give serious attention to the future structure of scientific organization and research in India not only within government but for the country as a whole.

11.7. If I may put it in this way, the solution does not lie in officialising all scientific activities but gradually to make the scientific organizations truly functional by transferring to them real responsibilities and entrusting them with fruitful activities. The advancement of science requires proper recognition of scientists in the structure of society as a whole. In this connexion I may remind you that Sir Alexander Fleck has given a quotation from an administrator which went on record that "the place of the expert is not in the saddle but as the assistant and the adviser. The scientist must be on tap, not on top." I can only say that the future of India depends on reversing the dictum of the administrator and recognising that the scientist must be on top and not merely on tap. I have a deep conviction that scientist must supply the leadership in national development. It is necessary that they should take the initiative.

Appendix

Statistical tables

TABLE 1. SOME CHARACTERISTICS OF DEVELOPMENT FOR SELECTED COUNTRIES

country	year	population in million (10^6)	steel production in million metric tons per year	per capita national income US dollars	physicians per ten thousand of population	school children percent of population
(1)	(2)	(3)	(4)	(5)	(6)	(7)
level of one million ton of steel per year						
1. USA	1880	50	1.3	635	16.3	19
2. UK	1880	31	1.3	385	6.6	12
3. Germany	1885	48	1.2	340	3.6	15
4. France	1895	39	0.9	485	3.9	14
5. Russia	1895	105	0.9	95	1.6	3
6. Japan	1923	56	1.0	125	7.7	17
7. India	1955–56	384	1.3	60	1.7	0.8
level of 4 or 5 million tons of steel per year						
1. USA	1890	63	4.3	735	15.9	20
2. UK	1900	38	5.0	560	8.8	16
3. Germany	1895	54	4.0	430	4.0	17
4. France	1913	39	4.7	550	5.3	14
5. USSR	1929	147	4.9	95	4.3	8
6. Japan	1935	69	4.9	195	8.4	17
7. China	1956	608	4.5	70	1.2	1.6
8. India	1960–61	408	4.3	72	2.1	1.0

Compiled by the Planning Division, Indian Statistical Institute.

Notes on Table 1

1. Figures of steel output are taken from Woytinsky and Woytinsky: *World Population and Production*, for all countries except China and India. Chinese and Indian figures relate to finished steel and are respectively taken from *China Today*, 25 August, 1957, and *Second Five Year Plan*, 1956 (India).

2. For USA, UK and Japan, population figures are taken from the Statistical Abstracts of the respective countries relating to different years. Figures for Germany and France are obtained by interpolation from data given by Woytinsky and Woytinsky. Russian data relating to 1895 are obtained in the same manner while the figure for 1926 is from the *National Economy of USSR*. Chinese data are from unpublished official Chinese sources; and Indian data from the *Second Five Year Plan*.

3. Data on number of school children for USA, Germany, UK and Russia are taken from the *Encyclopaedia of Social Reform* and Mulhall, *Dictionary of Statistics*. French data relating to 1895 is from the *Statistical Year Book*: *National Institute of Statistical and Economic Studies*, France, while for 1912-13, are from the *International Year Book of Statistics*, 1920. Data for USSR and India are from sources given in (2). The figure for Japan is based on the *Statistical Abstract of Japan, 1933* and the *UN Statistical Year Book*. The Chinese figure is from the *Report on Fulfilment of the National Economic Plan in 1955*.

4. The US figures for doctors are from the *Historical Statistics of USA* while those for UK are from the *Professions* by Carr-Saunders and Wilson. French data are obtained from Woytinsky and Woytinsky. The German and Russian figures are from Mulhall, *Dictionary of Statistics* and Woytinsky and Woytinsky, *World Population and Production* while the figure for USSR is from *40 years of Soviet Power*. The Japanese figures are from the *Statistical Abstract of Japan, 1933* and the *UN Statistical Year Book, 1949–50*. Indian and Chinese data are from sources indicated in (3).

5. National income figures are roughly dimensional and are obtained from certain estimates at current dollars for 1955 or 1956 carried backward by index numbers of real per capita income. Some of these current estimates were considered in the Science and National Planning, *Sankhyā*, September, 1958.

The index numbers used for USA are taken from data given in the *Economic Report of the President, 1957*; and the "Long-term Change in the National Income of the USA since 1870" by Simon Kuznets, (*Income and Wealth*, Series II). The German figure is adjusted by series available in the *Statistics of National Income and Expenditure*, (UN Statistical Papers, Series H), and the "Long-term Growth of National Income in Germany" by P. Jostock, (*Income and Wealth*, Series V). For the UK estimates, use has been made of data given by Prest, National Income of the UK, (*Economic Journal*, March 1948); Phyllis Dean, Contemporary Estimates of National Income......Nineteenth Century (*Economic History Review*, Vol. 8, No. 3 and Vol. 9, No. 3); the *Statistics of National Income and Expenditure* (*UN*) and *UN Statistical Year Book*. The French figures are obtained on the basis of UN sources indicated above and a series of real output per working person given in Colin Clark: *Conditions of Economic Progress*. The figure for Japan is adjusted by data given in the *Japanese Statistical Year Book*, 1957 while Chinese and Indian data relating to current period are taken from official sources.

Notes on Table 2

List A (developed countries) :

(1) Australia, (2) Austria, (3) Belgium, (4) Canada, (5) Denmark, (6) France, (7) Federal Republic of Germany, (8) Finland, (9) Iceland, (10) Italy, (11) Luxembourg, (12) Netherlands, (13) New Zealand, (14) Norway, (15) Sweden, (16) Switzerland, (17) South Africa, (18) UK, (19) USA.

List B (socialized countries) :

(1) Albania, (2) Bulgaria, (3) Czechoslovakia, (4) China, (5) German Democratic Republic, (6) Hungary, (7) Poland, (8) Rumania, (9) USSR, (10) Yugoslavia.

List C (medium developed countries) :

(1) Argentina, (2) Bolivia, (3) Brazil, (4) Costa Rica, (5) Cuba, (6) Chile, (7) Columbia, (8) Dominican Republic, (9) El-salvador, (10) Ecuador, (11) Greece, (12) Guatemala, (13) Haiti, (14) Honduras, (15) Ireland, (16) Japan, (17) Jamaica, (18) Mexico, (19) Portugal, (20) Paraguay, (21) Peru, (22) Panama, (23) Puerto Rico, (24) Spain, (25) Turkey, (26) Uruguay, (27) Venezuela.

List D (underdeveloped countries) :

(1) Afghanistan, (2) Algeria, (3) Angola, (4) Burma, (5) Belgian Congo, (6) China, (7) Ceylon, (9) Cameroon, (9) Egypt, (10) Ethiopea, (11) French West Africa, (12) Ghana, (13) Hongkong, (14) India, (15) Indonesia, (16) Iran, (17) Iraq, (18) Israel, (19) Kenya, (20) Lebanon, (21) Liberia, (22) Malay, (23) Morocco, (24) Mauritious, (25) Mozambique, (26) Madagascar, (27) Nigeria, (28) Pakistan, (29) Philippines, (30) Rhodesia and Nyasaland, (31) Saudi Arabia, (32) South Korea, (33) Syria, (34) Thailand, (35) Taiwan, (36) Tunisia, (37) Tanganayika, (38) Uganda, (39) Vietnam, (40) Yemen.

Sources :

Columns (2), (4), (5), (9) and (10) are from the *UN Statistical Year Book, 1956*.

The Chinese population figure is taken from statistics prepared by the State Statistical Bureau of the People's Republic of China (in Chinese). The Chinese figures for physicians and primary school students are from the statistics prepared by the State Statistical Bureau of the People's Republic of China.

The USSR figures for physicians and primary school students are from the data supplied by the Central Board of Statistics of the USSR.

The method of estimation of national income figures is explained generally in *Sankhyā*, Vol. 20 Parts 1 and 2, "Science and National Planning".

TABLE 2. POPULATION, NATIONAL INCOME, PRODUCTION OF STEEL AND ELECTRICITY, AND NUMBER OF PHYSICIANS AND PRIMARY SCHOOL STUDENTS BY COUNTRIES AND GROUPS OF COUNTRIES ABOUT 1955-56

countries	population in millions (10^6)	national income in US dollar billion (10^9)	production		per capita			physicians per ten thousand of population	primary school pupils per thousand of population
			crude steel in million (10^6) metric tons	electricity billion (10^9) kwh	national income in US dollar	production			
						crude steel (kg)	electricity (kwh)		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. United States of America	165	341	106	629	2030	642	3812	13.2	158
2. United Kingdom	50	46	20	94	900	394	1846	16.7	108
3. Developed (sub-total)[1]	450	566	191	1107	1292	436	2527	11.5	132
4. USSR	200	200	45	170	1000	226	850	14.9	65
5. China	620	44	3	12	72	4.8	20	1.2	87
6. Socialized (sub-total)[2]	900	278	63	298	317	72	340	5.9	90
7. Japan	90	21	10	65	230	105	724	10.5	144
8. Medium (sub-total)[3]	350	86	13	120	259	38	360	7.3	116
9. India	380	23	2	8	61	4.5	21	1.8	54
10. Underdeveloped (sub-total) excluding China	900	71	2	25	80	2	28	1.4	65
11. Underdeveloped including China[4]	1500	114	5	37	76	3	25	1.3	74
12. World [3+6+8+10] total	2600	1002	269	1550	395	106	611	5.5	92

Compiled by the Planning Division, Indian Statistical Institute. Population and other figures have been rounded off dimensionally.

[1] Includes developed countries in North America, Europe, Australia (as in list A).

[2] Includes all socialized countries (as in list B).

[3] Includes Japan and medium developed countries in Europe and Latin America (as in list C).

[4] Includes underdeveloped countries in Asia and Africa (as in list D).

TABLE 3. OUTPUT OF CRUDE STEEL IN MILLION METRIC TONS[1] FOR WORLD AND SELECTED COUNTRIES

year (1)	World[2] (2)	United Kingdom (3)	United States (4)	Germany (5)	USSR (6)	Japan[6] (7)	China (8)	India (9)
1870	0.7	0.2	f[3]	0.2	f			
1880	4.4	1.3	1.3	0.7	0.3			
1890	12.4	3.6	4.3	2.2	0.4			
1900	28.3	5.0	10.4	6.6	2.2			
1910	60.5	6.5	26.5	13.7	3.5			
1913	76.5	7.8	31.8	18.9	4.8	0.2		f
1918	78.6	9.7	45.2	15.0	0.4	0.8		0.2
1923	78.5	8.6	45.7	6.3	0.6	1.0		0.2
1929	120.5	9.8	57.3	16.2	4.9	2.3		0.6
1932	50.7	5.3	13.9	5.7	5.9	2.4		0.6[10]
1939	135.6	13.4	47.9	22.4	17.6	6.7	0.5	1.1
1940	142.9	13.2	60.8	19.0	18.3	6.9	0.6	1.3
1944	152.1	12.3	81.3	18.0	13.9	5.9	0.9[7]	1.4
1945	113.3	12.0	72.3	0.3	12.3	2.1	n.a.	1.4
1950	188.7	16.6	87.8	12.1[4]	27.3	4.8	0.2[8]	1.5
1955	269.1	20.1	106.2	21.3	45.3	9.4	2.9	1.7
1956	282.6	21.0	104.5	23.2	48.6[5]	11.1	4.5[9]	1.8

[1] one metric ton = 1.1023 short tons.

[2] Excludes China upto 1932 and 1944-50.

[3] f = less than 100,000 tons.

[4] West Germany from 1950.

[5] USSR 1957 and 1958 figures are 51 and 55 million metric tons respectively.

[6] Includes Korea upto 1932.

[7] Relates to 1943.

[8] Relates to 1949.

[9] China 1957 and 1958 figures are 5.3 and 11 million metric tons respectively.

[10] Relates to 1932-33.

Compiled by the Planning Division, Indian Statistical Institute, from :

(*a*) Woytinsky and Woytinsky : *World Population and Production : Trends and Outlook* (New York, 1953).

(*b*) *United Nations Statistical Year Book, 1957.*

(*c*) D. G. Zhimerin : *Economy of the Soviet Union, Past and Present,* (Moscow, 1958).

(*d*) *News and Views from the Soviet Union,* (USSR Embassy, New Delhi, 18 December 1958).

(*e*) *China Today :* 25 September 1958; 10 January 1959.

(*f*) *Report on Fulfilment of the National Economic Plan of the People's Republic of China in 1955* (with Statistical Summary).

(*g*) Sastry : *A Statistical Study of India's Industrial Development.*

(*h*) *Statistical Abstract,* India : 1927-28 to 1936-37.

TABLE 4. INCREASE IN LIVE NUMBER ON LABOUR EXCHANGE REGISTERS IN INDIA IN 5 YEARS: 1953–1957

occupational category	number on live-register at end of month		increase in five years		monthly average of effective placements		ratio of applicants to effective placements[1]	
	January 1953 (131)	December 1957 (181)	number	percent	1953	1957		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. industrial supervisory	4,317	5,929	1,612	37.3	180	171	24.0	34.7
2. skilled and semiskilled	46,539	71,508	24,969	53.7	2,247	1,910	20.7	37.4
3. educational	6,625	40,246	33,621	507.5	401	1,117	16.5	36.0
4. clerical	113,424	267,757	154,333	136.1	2,006	3,594	56.5	74.5
5. domestic service	13,987	32,018	18,031	128.9	938	1,141	14.9	28.1
6. unskilled	221,761	460,639	238,878	107.7	8,452	6,178	26.2	74.6
7. others	27,775	44,002	16,227	58.4	1,230	1,958	22.6	22.5
8. all categories	434,428	922,099	487,671	112.3	15,454	16,069	28.1	57.4

[1] That is, col. (8) = col. (2)/col. (6); and col. (9) = col. (3)/col. (7).

TABLE 5. NUMBER ON LIVE REGISTER AND NUMBER OF VACANCIES AND REPLACEMENTS: LABOUR EXCHANGES IN INDIA, 1953–1957

all categories	year				
	1953	1954	1955	1956	1957
(1)	(2)	(3)	(4)	(5)	(6)
1. monthly average of effective placements	15,454	13,538	14,145	15,821	16,069
2. monthly average of vacancies notified	21,392	19,990	23,377	24,718	24,766
3. monthly average on Live Register	477,575	562,319	647,109	749,099	816,280
4. number of applicants per vacancy (row 3 ÷ row 2)	22.3	28.1	27.7	30.3	33.0
5. number of applicants per placement (row 3 ÷ row 1)	30.9	41.5	45.7	47.3	50.8

Compiled in the Indian Statistical Institute from "A Study of Trends in the Number of Employment Seekers in the Year 1953–1957." (Ministry of Labour & Employment, D.G.R.& E., September 1958).

TABLE 6. DISTRIBUTION OF DOCTORS PER MILLION POPULATION BY LEVEL OF QUALIFICATIONS AND STATE OF RESIDENCE IN 1954

states	population 1954 (million)	number of doctors			doctors per million (10^6) population			index with West Bengal = 100
		graduate and post-graduate	licentiate	all	graduate and post-graduate	licentiate	all	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. Andhra	32.65	2,025	1,908	3,933	62	58	120	19.8
2. Assam	9.59	403	2,281	2,684	42	238	280	46.3
3. Bihar	40.07	1,762	3,302	5,064	44	82	126	20.8
4. Bombay	51.15	6,928	6,044	12,972	135	118	253	41.8
5. Jammu & Kashmir	4.54	32	42	74	7	9	16	2.6
6. Kerala	14.51	836	845	1,681	58	58	116	19.2
7. Madhya Pradesh	26.80	731	1,191	1,922	28	44	72	11.9
8. Madras	31.38	2,959	3,083	6,042	95	98	193	31.9
9. Mysore	20.58	1,100	1,846	2,946	53	90	143	23.6
10. Orissa	14.94	448	800	1,248	30	54	84	13.9
11. Punjab	16.63	1,118	1,915	3,033	67	115	182	30.1
12. Rajasthan	16.74	213	454	667	13	27	40	6.6
13. Uttar Pradesh	65.60	3,244	2,989	6,233	49	46	95	15.7
14. West Bengal	27.41	6,157	10,434	16,591	225	381	605	100.0
15. Delhi	1.93	771	468	1,239	399	242	642	106.1
16. Himachal Pradesh	1.13	31	42	73	27	37	65	10.7
17. Manipur	0.60	3	21	24	5	35	40	6.6
18. Tripura	0.69	31	145	176	45	210	255	42.1
all India	376.94	28,792	37,810	66,602	76	100	176	29.1

Source : Perspective Planning Division, Planning Commission; based on Medical Registers.

TABLE 6E. INCOME AND EXPENDITURE OF UNIVERSITIES AND COLLEGES IN INDIA : 1955–56

type of instituion	recurring expenditure (Rs. lakhs)	source of income (percentage)				type of expenditure (percentage)		
		govt. grants	fees	endowment	others	salary	scholarship	others
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. university teaching	937	49	34	2	15	25	4	71
2. boards of education	132	1	97	—	2	—	—	100
3. professional colleges	700	67	23	3	7	48	14	38
4. general colleges	1040	34	56	3	7	62	14	24
5. intermediate colleges	125	39	39	7	15	68	22	9
6. total higher	2934	45	42	3	10	44	10	46
in Rs. lakhs	—	1320	1232			1291	293	

Source : Education in India, Vol. 1 1955-56.

TABLE 7. STRUCTURE OF EDUCATION: SELECTED COUNTRIES

level of education	USA 1955	USSR 1957–58	UK 1954	West Germany 1956	France 1954	Japan 1956	China 1955	India 1955–56
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1. population (in millions)	165	200	51	51	43	90	608	384
number of students in thousands (000's)								
2. primary	25,458	25,055	5,484	4,955	4,921	12,956	53,126	22,899
3. secondary	7,961	7,511	2,483	3,668	1,300	8,063	4,473	8,789
4. higher	2,679	2,100	84	127	150	610	292	681
5. total	36,098	34,666	8,051	8,750	6,371	21,629	57,891	32,369
number of students per 1000 of population								
6. primary	154.3	125.3	107.5	97.2	114.0	144.4	87.4	59.6
7. secondary	48.2	37.6	48.7	71.9	30.2	89.6	7.4	22.9
8. higher	16.2	10.5	1.6	2.5	3.5	6.8	0.5	1.8
9. total	218.7	173.4	157.8	171.6	148.1	240.4	95.3	84.3
number of secondary and higher students per 1000 primary students								
10. secondary	312.7	299.8	452.8	740.3	264.2	622.3	84.2	383.8
11. higher	105.2	83.8	15.3	25.6	30.5	47.1	5.5	29.7
12. total	417.9	383.6	468.1	765.9	294.7	669.4	89.7	413.5
number of higher students per 1000 secondary students								
13. higher	336.5	279.6	33.8	34.6	115.4	75.7	65.3	77.5

Compiled by the Planning Division, Indian Statistical Institute, from:

1. Population figures are taken from *UN Statistical Year Book* 1956 and 1957 except for China which is taken from official data (in Chinese).

2. Educational statistics are taken from the following sources:

USA: *Statistical Abstract of USA*, 1956 and 1957.

USSR: *USSR in Figures* (in Russian), 1958. Structural features of Soviet education are not fully revealed in recent data due to the fact that very few children were born during the war.

UK, West Germany, France and Japan: *UN Statistical Year Book*, 1957.

China: From *Report on Fulfilment of the National Economic Plan of the People's Republic of China in* 1955.

India: Basic Statistics relating to Indian Economy (mimeographed: CSO) for primary education; and *Education in India*, Vol. 1, 1955–56.

TABLE 8. ENROLMENT OF STUDENTS BY LEVEL, SUBJECT AND SEX WITH PERCENTAGE : 1950–51 and 1956–57

level and subject	1950–51						1956–57						index 1950–51 =100
	number (000)			percentage			number (000)			percentage			
	boys	girls	total	boys	girls	total	boys	girls	total	boys	girls	total	
(1)	(2.1)	(3.1)	(4.1)	(2.2)	(3.2)	(4.2)	(5.1)	(6.1)	(7.1)	(5.2)	(6.2)	(7.2)	(8)
1. *Intermediate*	198.6	24.4	223.0	89.1	10.9	100.0	381.2	59.5	440.7	86.5	13.5	100.0	198
2. arts	96.3	18.0	114.3	43.2	8.0	51.2	195.1	48.1	243.2	44.3	10.9	55.2	213
3. commerce	16.1	0.1	16.2	7.3	0.0	7.3	35.0	0.2	35.2	7.9	0.1	8.0	217
4. science	84.6	6.3	90.9	37.9	2.9	40.8	145.0	11.2	156.2	32.9	2.5	35.4	172
5. agriculture	1.6	—	1.6	0.7	—	0.7	6.1	—	6.1	1.4	—	1.4	381
6. *Diploma*	9.5	0.3	9.8	97.0	3.0	100.0	26.1	0.7	26.8	97.4	2.6	100.0	273
7. ,, engineering and technology	8.2	—	8.2	83.7	—	83.7	23.7	—	23.7	88.4	—	88.4	289
8. ,, education	0.4	0.2	0.6	4.1	2.0	6.1	0.7	0.5	1.2	2.6	1.9	4.5	200
9. ,, medicine and veterinary	0.9	0.1	1.0	9.2	1.0	10.2	1.7	0.2	1.9	6.4	0.7	7.1	190
10. *Degree*	115.1	13.7	128.8	89.4	10.6	100.0	210.2	34.7	244.9	85.8	14.2	100.0	190
11. arts	44.4	8.9	53.3	34.5	6.9	41.4	88.8	25.0	113.8	36.3	10.2	46.5	214
12. commerce	15.8	0.1	15.9	12.3	—	12.3	31.1	0.2	31.3	12.7	0.1	12.8	197
23. science	29.1	2.4	31.5	22.6	1.9	24.5	48.6	5.8	54.4	19.8	2.4	22.2	173
14. engineering & technology[1]	10.1	—	10.1	7.8	—	7.8	17.7	—	17.7	7.2	—	7.2	175
15. medicine & veterinary[2]	13.0	2.3	15.3	10.1	1.8	11.9	20.3	3.7	24.0	8.3	1.5	9.8	157
16. agriculture	2.7	—	2.7	2.1	—	2.1	3.7	—	3.7	1.5	—	1.5	137
17. *Post-graduate*	34.0	3.9	37.9	89.7	10.3	100.0	56.9	9.3	66.2	86.0	14.0	100.0	175
18. arts	11.0	2.1	13.1	29.0	5.5	34.5	17.3	4.2	21.5	26.1	6.3	32.4	164
19. commerce	2.0	—	2.0	5.3	—	5.3	2.7	—	2.7	4.1	—	4.1	135
20.1. law	13.2	0.3	13.5	34.8	0.8	35.6	18.9	0.4	19.3	28.6	0.6	29.2	143
20.2. law (higher)	0.2	—	0.2	0.5	—	0.5	0.4	—	0.4	0.6	—	0.6	200
21.1. education (B.T.)	2.1	1.0	3.1	5.6	2.6	8.2	7.6	3.3	10.9	11.5	5.0	16.5	352
21.2. education (higher)	0.3	0.1	0.4	0.8	0.3	1.1	0.5	0.3	0.8	0.8	0.4	1.2	200
22. science	4.3	0.4	4.7	11.3	1.1	12.4	7.3	0.9	8.2	11.0	1.4	12.4	174
23.1. technology & engineering[1]	0.2	—	0.2	0.5	—	0.5	0.6	—	0.6	0.9	—	0.9	300
23.2. medicine & veterinary[2]	0.3	—	0.3	0.8	—	0.8	1.0	0.2	1.2	1.5	0.3	1.8	400
24. agriculture	0.4	—	0.4	1.1	—	1.1	0.6	—	0.6	0.9	—	0.9	150
Intermediate													
25. arts, commerce	112.4	18.1	130.5	50.5	8.0	58.5	230.1	48.3	278.4	52.2	11.0	63.2	213
26. science & technology	86.2	6.3	92.5	38.6	2.9	41.5	151.1	11.2	162.3	34.3	2.5	36.8	175
Degree												100.0	
27. arts, commerce	60.2	9.0	69.2	46.8	6.9	53.7	119.9	25.2	145.1	49.0	10.3	59.3	210
28. science & technology	54.9	4.7	59.6	42.6	3.7	46.3	90.3	9.5	99.8	36.8	3.9	40.7	167
Post-graduate												100.0	
29. arts, law & commerce	26.4	2.4	28.8	69.6	6.3	75.9	39.3	4.6	43.9	59.2	7.1	66.3	152
30. science & technology[2]	5.2	0.4	5.6	13.7	1.1	14.8	9.5	1.1	10.6	14.3	1.7	16.0	189
31. education	2.4	1.1	3.5	6.4	2.9	9.3	8.1	3.6	11.7	12.3	5.4	17.7	334
32. *All-levels*	357.2	42.3	399.5	89.4	10.6	100.0	674.4	104.2	778.6	86.6	13.4	100.0	195
35. arts, law, commerce	199.0	29.5	228.5	49.8	7.4	57.2	389.9	78.1	468.0	50.1	10.0	60.1	205
34. education[3]	2.8	1.3	4.1	0.7	0.3	1.0	8.8	4.1	12.9	1.1	0.6	1.7	315
35. science & technology[2]	155.4	11.5	166.9	38.9	2.9	41.8	275.7	22.0	297.7	35.4	2.8	38.2	178

---

[1] Includes agriculture. [2] Includes engineering, medicine and veterinary, agriculture. [3] Includes diploma in education.

ource : Perspective Planning Division, Planning Commission.

TABLE 9. OUTTURN OF QUALIFIED PERSONS IN INDIA BY LEVEL AND SUBJECT : 1950-51 AND 1955-56

examination	1950-51				1955-56				index 1950-51 =100
	number passed			percentage of girls	number passed			percentage of girls	
	boys	girls	total		boys	girls	total		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. *Intermediate*	71,272	9,815	81,087	12.1	125,939	20,213	146,152	13.8	180
2. arts	38,748	7,913	46,661	17.0	75,764	16,482	92,246	17.9	198
3. commerce	6,393	17	6,410	0.3	11,216	104	11,320	0.9	177
4. science	25,459	1,884	27,343	6.9	37,716	3,625	41,341	8.8	151
5. agriculture	672	1	673	0.1	1,243	2	1,245	0.2	185
6. *Diploma*	8,724	5,418	14,142	38.3	10,031	2,535	12,566	20.2	89
7. ,, engineering & technology	2,026	1	2,027	0.1	4,977	—	4,977	—	246
8. ,, others	6,698	5,417	12,115	44.7	5,054	2,535	7,589	33.4	63
9. *Degree*	38,571	5,380	43,951	12.2	60,912	10,648	71,560	14.9	163
10.1. arts (pass)	16,586	3,738	20,324	18.4	29,117	7,899	37,016	21.3	182
10.2. ,, (hons.)	1,814	450	2,264	19.9	1,865	487	2,352	20.7	104
11. commerce	5,411	22	5,433	0.4	8,203	50	8,253	0.6	152
12.1. science (pass)	9,640	767	10,407	7.4	13,370	1,577	14,947	10.6	144
12.2. ,, (hons.)	713	73	786	9.3	1,077	132	1,209	10.9	154
13. engineering & technology	1,772	1	1,773	0.1	3,810	6	3,816	0.2	215
14. medicine & veterinary	1,598	325	1,923	16.9	2,583	491	3,074	16.0	160
15. agriculture	1,037	4	1,041	0.4	887	6	893	0.7	86
16. *Post-graduate*	12,993	2,052	15,045	13.6	23,527	5,038	28,565	17.6	190
17. arts	5,020	949	5,969	15.9	7,627	1,901	9,528	20.0	160
18. commerce	575	1	576	0.2	1,027	4	1,031	0.4	179
19.1. law	3,474	71	3,545	2.0	5,733	118	5,851	2.0	165
19.2. law (higher)	15	1	16	6.2	41	2	43	4.7	269
20.1. education (B.T.)	2,096	881	2,977	29.6	6,052	2,576	8,628	29.9	290
20.2. education (higher)	151	65	216	30.1	301	142	443	32.1	205
21. science	1,326	72	1,398	5.2	2,277	262	2,539	10.3	182
22.1. engineering	87	1	88	1.1	148	—	148	—	168
22.2, medicine & veterinary	106	11	117	9.4	169	33	202	16.3	173
23. agriculture	143	—	143	—	152	—	152	—	106
Intermediate									
24. arts & commerce	45,141	7,930	53,071	14.9	86,980	16,586	103,566	16.0	195
25. science, agriculture & technology	26,131	1,885	28,016	6.7	38,959	3,627	42,586	8.5	152
Degree									
26. arts, law & commerce	23,811	4,210	28,021	15.0	39,185	8,436	47,621	17.7	170
27. science & technology[1]	14,760	1,170	15,930	7.3	21,727	2,212	23,939	9.2	150
Post-graduate									
28. arts, Law & commerce	9,084	1,022	10,106	10.1	14,428	2,025	16,453	12.3	163
29. science & technology[1]	1,662	84	1,746	4.8	2,746	295	3,041	9.7	174
30. *All levels*	131,560	22,665	154,225	14.7	220,409	38,434	258,843	14.8	168
31. arts, law & ommerce	78,036	13,162	91,198	14.4	140,593	27,047	167,640	16.1	184
32. education	2,247	946	3,193	29.6	6,353	2,718	9,071	30.0	284
33. science & technology[1]	44,579	3.140	47.719	6.6	68,409	6,134	74,543	8.2	156
34. doctorate	115	13	128	10.2	261	24	285	8.4	223

[1] Includes engineering, medicine & veterinary, and agriculture.

Source : Perspective Planning Division, Planning Commission.

TABLE 10 NUMBER OF STUDENTS APPEARING IN AND PASSING DIFFERENT EXAMINATIONS, AND THE PERCENTAGE OF SUCCESSFUL STUDENTS

India : 1951–1952 to 1955–1956

	1951–1952	1952–1953	1953–1954	1954–1955	1955–1956
(1)	(2)	(3)	(4)	(5)	(6)
Appeared					
1. Intermediate Arts	109,493	139,422	169,287	189,385	207,118
2. Intermediate Science	61,891	67,155	78,973	85,042	93,405
3. Bachelor of Arts	47,841	54,710	74,824	81,421	79,920
4. Bachelor of Science	23,428	27,457	31,107	34,754	32,667
5. Master of Arts	7,780	7,808	10,281	11,729	13,215
6. Master of Science	2,144	2,114	2,667	2,946	3,142
7. Professional	36,402	41,497	48,237	53,005	48,450
8. total	288,979	340,163	415,376	458,282	477,917
Passed					
9. Matriculation	261,059	334,760	397,005	400,014	429,494
10. Intermediate Arts	52,373	59,703	71,640	86,741	90,182
11. Intermediate Science	25,463	29,318	33,211	39,735	41,557
12. Bachelor of Arts	24,965	27,491	35,773	40,445	37,992
13. Bachelor of Science	11,171	12,526	14,405	16,704	15,997
14. Master of Arts	6,063	6,169	7,753	8,892	9,313
15. Master of Science	1,680	1,694	2,068	2,211	2,456
16. Professional	22,684	26,269	30,162	33,181	35,772
16.1. Commerce	6,785	7,759	8,452	8,773	9,597
16.2. Law	4,538	5,562	6,602	5,996	5,628
16.3. Education & Teacher's Training[1]	5,251	6,215	6,749	9,492	11,106
16.4. Technology & Engineering	2,592	3,117	3,638	3,680	4,444
16.5. Medicine & Veterinary	2,324	2,512	3,541	4,089	3,888
16.6. Agriculture and Forestry	1,194	1,104	1,180	1,151	1,109
17. total	405,458	407,930	592,017	627,923	662,763
Percentage passed					
18. Intermediate Arts	47.8	42.8	42.3	45.8	43.5
19. Intermediate Science	41.1	43.7	42.1	46.7	44.5
20. Bachelor of Arts	52.2	50.2	47.8	49.6	47.5
21. Bachelor of Science	47.7	45.6	46.3	48.1	49.0
22. Master of Arts	77.9	79.0	75.4	75.8	70.5
23. Master of Science	78.4	80.1	77.5	75.1	78.2
24. Intermediate (I.A. & I.Sc.)	45.4	43.1	42.2	46.1	43.8
25. Degree (B.A. & B.Sc.)	50.7	48.7	47.4	49.2	48.0
26. M.A. & M.Sc.	78.0	79.2	75.8	75.7	72.0
27. Professional	62.3	63.3	62.5	62.6	73.8
28. total	50.0	48.0	46.9	49.7	48.8

---

[1] Includes "others."

Source : *Education in India* 1952–1953, 1954–1955 and 1955–1956.

TABLE 11. NUMBER AND PROPORTION OF STUDENTS IN HIGHER COURSES IN RELATION TO NUMBER OF MATRICULATES IN INDIA IN 1952

courses	number on rolls			per cent			number per 10,000 matriculates		
	boys	girls	total	boys	girls	total	boys	girls	total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Intermediate	first year intermediate courses in 1952-53								
(0.1) number of matriculates in 1952	222,371	35,767	258,138	100	100	100	10,000	10,000	10,000
1. total on rolls (1952-53)	153,941	20,330	174,271	69.2	56.8	67.5	6,920	5,680	6,750
2. arts	63,851	13,621	77,472	28.7	38.1	30.0	2,870	3,810	3,000
3. commerce	19,510	127	19,637	8.8	0.4	7.6	880	40	760
4. science	62,548	5,903	68,451	28.2	16.4	26.5	2,820	1,640	2,650
5. agriculture	1,209	—	1,209	0.5	—	0.5	50	—	50
6. diploma : technology	5,667	1	5,668	2.5	—	2.2	250	—	220
7.1. ,, : education	235	201	436	0.1	0.6	0.2	10	60	20
7.2. ,, : others	921	477	1,398	0.4	1.3	0.5	40	130	50
Degree	first year degree courses in 1954-55								
(0.2) number passed intermediate in 1954	98,011	15,612	113,623	100	100	100	4,408	4,365	4,402
8. total on rolls (1954-55)	74,947	10,727	85,674	76.5	68.7	75.4	3,370	2,999	3,319
9. arts (pass & hons.)	33,262	7,654	40,916	33.9	49.1	36.0	1,496	2,140	1,585
10. commerce	12,515	77	12,592	12.8	0.5	11.1	563	22	488
11. science	19,567	2,146	21,713	20.0	13.7	19.4	880	600	841
12. technology & engineering	4,000	1	4,001	4.1	—	3.5	180	—	155
13. medicine & veterinary	3,056	733	3,789	3.1	4.7	3.3	137	205	147
14. agriculture	1,221	5	1,226	1.2	—	1.8	55	1	47
15. diploma (others)	1,326	111	1,437	1.4	0.7	1.3	59	31	56
Post-graduate	first year post-graduate courses in 1956-57								
(0.3) numbers awarded degrees in 1956	51,698	9,332	61,030	100	100	100	2,325	2,609	2,364
16. total on rolls (1956-57)	33,988	6,084	40,072	65.7	65.2	65.7	1,528	1,701	1,552
17. arts	9,015	2,023	11,038	17.4	21.7	18.1	405	566	428
18. commerce	1,422	11	1,433	2.8	0.1	2.3	64	3	55
19. law	11,698	220	11,918	22.6	2.4	19.5	526	61	462
20. education	7,349	3,295	10,644	14.2	35.3	17.5	330	921	412
21. science	3,375	382	3,757	6.5	4.1	6.2	152	107	146
22. agriculture	264	3	267	0.5	—	0.4	12	1	10
23. diploma (others)	865	150	1,015	1.7	1.6	1.7	39	42	39
Science & technology	first year science & technology (totals)								
24. intermediate (1952-53)	69,424	5,904	75,328	31.2	16.4	29.2	3,120	1,640	2,920
25. degree (1954-55)	27,844	2,885	30,729	28.4	18.4	28.0	1,252	806	1,190
26. post-graduate (1956-57)	3,639	385	4,024	7.0	4.1	6.6	164	108	156

Source : Perspective Planning Division, Planning Commission.

Note : Percentage figures given in cols. (5), (6) and (7) are on the basis of the number of candidates who passed the appropriate previous examination as given in lines (0.1), (0.2) and (0.3).

TABLE 12. TOTAL CONSUMER EXPENDITURE PER PERSON PER 30 DAYS AND CUMULATIVE PERCENTAGES OF POPULATION AND CONSUMER EXPENDITURE BY SUB-SAMPLES

National Sample Survey : 9th Round : May—November 1955 ; All-India (Rural)[1]

per-centile	number of sample households			total consumer expenditure in Rs. per person per 30 days						cumulative percentage of					
				limiting values			average expenditure			total population			total consumer expenditure		
	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
5	39	39	78	4.91	4.79	4.90	4.01	3.82	3.90	6.32	5.58	5.83	1.69	1.39	1.50
10	38	35	71	5.99	5.85	5.92	5.43	5.39	5.39	12.39	10.86	11.56	3.88	3.26	3.54
15	35	37	73	6.74	6.88	6.80	6.39	6.38	6.38	17.17	15.94	16.50	5.91	5.37	5.62
20	36	43	78	7.54	7.90	7.76	7.25	7.35	7.28	22.16	21.47	21.72	8.32	8.03	8.13
25	37	37	76	8.52	8.59	8.56	8.05	8.31	8.20	26.82	27.05	26.98	10.81	11.06	10.97
30	33	42	75	9.36	9.57	9.53	8.96	9.19	9.06	32.78	32.49	32.61	14.36	14.33	14.34
35	38	41	79	10.17	10.39	10.31	9.80	9.95	9.89	37.43	38.06	37.78	17.39	17.95	17.71
40	38	34	72	11.03	11.09	11.07	10.66	10.74	10.69	42.93	43.75	43.13	21.29	21.95	21.48
45	34	36	73	11.80	12.06	11.92	11.42	11.68	11.49	47.56	48.50	47.74	24.80	25.58	24.97
50	41	43	80	12.83	13.02	12.88	12.35	12.59	12.43	53.11	53.41	53.12	29.36	29.62	29.38
55	40	38	78	14.14	14.05	14.14	13.46	13.45	13.44	58.19	58.65	58.37	33.90	34.23	34.03
60	37	43	80	15.13	15.55	15.24	14.57	14.73	14.64	63.77	63.75	63.83	39.30	39.14	39.30
65	42	34	77	16.61	16.97	16.77	15.81	16.12	16.00	68.51	67.87	68.32	44.29	43.48	44.04
70	44	38	82	18.49	18.59	18.56	17.61	17.88	17.78	74.26	72.37	73.38	51.02	48.75	49.97
75	40	39	78	19.85	20.08	20.02	19.14	19.37	19.26	78.60	77.55	77.94	56.55	55.30	55.76
80	37	40	76	22.59	22.01	22.11	21.22	21.27	21.19	82.95	82.65	82.73	62.69	62.40	62.45
85	39	31	72	24.95	23.60	24.35	23.77	22.97	23.36	87.83	86.19	87.07	70.40	67.71	69.12
90	40	40	79	29.57	27.53	28.59	27.40	25.42	26.23	92.60	91.08	91.47	79.09	75.85	76.74
95	40	40	81	38.05	38.38	38.13	32.85	32.06	32.17	96.92	96.07	96.46	88.54	86.29	87.31
100	40	38	78	194.41	128.86	194.41	55.96	53.30	54.31	100	100	100	100	100	100
all	768	768	1536				15.03	15.24	15.15						

Number of sample villages including 4 uninhabited villages in each sub-sample : 772 (s.s.1), 772 (s.s.2), 1544 (combined).
Number of sample households : 768 (s.s.1), 768 (s.s.2), 1536 (combined).

[1] Excludes Jammu & Kashmir.

TABLE 13. EXPENDITURE IN RUPEES PER PERSON PER 30 DAYS BY ITEMS OF CONSUMPTION AND BY SUB-SAMPLES

National Sample Survey : 9th Round : May—November 1955 ; All-India (Rural)[1]

per-centile	population[2] in lakh (10^5)	food grains			food total			sugar			clothing			education			medicine			medical services		
		s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
5	197	2.34	2.49	2.39	3.12	3.18	3.12	0.04	0.04	0.04	0.10	0.05	0.08	0.00	0.02	0.01	0.01	0.00	0.01	0.00	–	0.00
10	194	3.01	3.14	3.09	4.41	4.22	4.32	0.06	0.06	0.06	0.09	0.23	0.16	–	0.01	0.01	0.06	0.04	0.05	–	–	–
15	167	3.43	3.54	3.48	5.05	4.94	5.02	0.10	0.11	0.11	0.09	0.42	0.22	0.02	0.01	0.02	0.03	0.04	0.04	0.02	0.02	0.02
20	177	3.42	3.88	3.60	5.15	5.63	5.32	0.13	0.11	0.12	0.33	0.59	0.52	0.00	0.01	0.00	0.07	0.01	0.04	0.05	–	0.02
25	178	4.46	4.37	4.50	6.43	6.48	6.47	0.09	0.21	0.14	0.32	0.34	0.31	0.00	0.05	0.03	0.03	0.01	0.02	–	–	–
30	191	4.28	4.50	4.39	6.73	6.86	6.83	0.25	0.16	0.23	0.27	0.61	0.44	0.02	0.01	0.02	0.08	0.03	0.06	0.01	0.06	0.04
35	175	5.11	5.06	5.05	7.43	7.81	7.56	0.28	0.26	0.24	0.72	0.72	0.74	0.03	0.00	0.01	0.03	0.02	0.02	0.04	–	0.02
40	181	5.44	5.43	5.34	8.16	8.11	8.11	0.26	0.25	0.28	0.61	0.66	0.67	0.08	0.02	0.05	0.05	0.01	0.04	–	0.01	–
45	156	5.18	6.02	5.61	8.59	9.00	8.79	0.35	0.28	0.30	0.66	0.63	0.62	0.04	0.02	0.04	0.15	0.23	0.15	0.02	–	0.02
50	182	6.25	5.70	6.00	9.40	9.65	9.52	0.16	0.34	0.28	0.69	0.57	0.62	0.08	0.04	0.06	0.11	0.08	0.12	–	–	0.01
55	177	5.45	6.09	5.80	9.38	9.97	9.70	0.38	0.35	0.34	1.24	0.79	0.99	0.22	0.14	0.17	0.11	0.09	0.10	0.04	0.02	0.03
60	185	7.05	5.99	6.56	11.13	10.22	10.88	0.33	0.44	0.43	1.00	0.97	1.02	0.05	0.21	0.11	0.15	0.33	0.17	0.08	0.00	0.04
65	152	6.29	6.09	6.30	11.01	11.28	11.12	0.64	0.66	0.60	1.85	1.38	1.62	0.04	0.01	0.05	0.21	0.18	0.23	0.08	0.16	0.08
70	171	5.68	7.03	6.18	12.07	12.22	11.98	0.51	0.60	0.53	1.74	1.67	1.69	0.13	0.15	0.13	0.18	0.19	0.24	0.02	0.02	0.06
75	154	7.15	7.23	7.19	13.32	13.44	13.36	0.49	0.66	0.58	2.19	1.88	2.04	0.12	0.07	0.11	0.18	0.32	0.24	0.02	0.05	0.04
80	162	7.41	7.61	7.50	14.22	14.40	14.30	0.87	0.49	0.69	2.12	2.50	2.27	0.21	0.09	0.14	0.18	0.52	0.39	0.08	0.07	0.07
85	147	7.82	8.49	8.05	15.44	16.70	16.25	0.73	0.71	0.76	2.11	0.92	1.50	0.23	0.20	0.22	0.81	0.46	0.49	0.26	0.18	0.14
90	149	8.36	8.09	8.21	16.28	16.52	16.37	0.68	0.98	0.85	3.43	2.53	2.89	0.15	0.29	0.30	0.82	0.44	0.82	0.19	0.19	0.29
95	169	9.66	7.21	8.50	20.13	17.80	18.30	1.14	1.27	1.08	3.86	4.27	4.24	0.71	0.51	0.46	0.43	1.32	0.84	–	0.39	0.21
100	120	10.51	8.97	9.50	22.88	22.65	23.02	0.98	1.24	1.19	5.87	4.90	5.28	0.58	0.29	0.48	8.37	1.12	4.21	3.43	0.42	1.69
all	3384	5.70	5.71	5.70	9.99	10.16	10.08	0.39	0.44	0.41	1.31	1.27	1.29	0.12	0.10	0.11	0.44	0.25	0.34	0.15	0.07	0.11

expenditure in rupees per person per 30 days (columns 3–23)

Number of sample villages including 4 uninhabited villages in each sub-sample : 772 (s.s.1), 772 (s.s.2), 1544 (combined).
Number of sample households : 768 (s.s.1), 768 (s.s.2), 1536 (combined).

[1] Excludes Jammu & Kashmir.

[2] Population estimates used elsewhere are based on projection.

TABLE 14. CUMULATIVE PERCENTAGES OF EXPENDITURE PER 30 DAYS BY ITEMS OF CONSUMPTION AND BY SUB-SAMPLES

National Sample Survey : 9th Round : May—November 1955 ; All-India (Rural)[1]

per-centile	cumulative percentage of expenditure on																				
	food grains			food total			sugar			clothing			education			medicine			medical services		
	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.	s.s.1	s.s.2	comb.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
5	2.59	2.43	2.44	1.98	1.75	1.80	0.61	0.51	0.52	0.48	0.20	0.34	0.20	0.90	0.52	0.15	0.06	0.12	0.07	—	0.05
10	5.80	5.33	5.54	4.65	3.94	4.26	1.56	1.25	1.34	0.88	1.16	1.04	0.20	1.48	0.79	0.99	0.84	0.94	0.07	—	0.05
15	8.68	8.48	8.55	7.07	6.41	6.72	2.76	2.51	2.62	1.22	2.83	1.88	0.98	1.99	1.45	1.35	1.60	1.45	0.66	1.45	0.92
20	11.67	12.24	11.85	9.64	9.47	9.48	4.39	3.86	4.10	2.47	5.42	4.01	1.14	2.25	1.54	2.12	1.76	2.01	2.38	1.45	2.08
25	15.31	16.51	16.00	12.64	13.03	12.85	5.42	6.57	5.90	3.63	6.93	5.27	1.17	4.87	2.88	2.45	1.91	2.26	2.38	1.45	2.08
30	19.79	20.80	20.34	16.65	16.70	16.67	9.19	8.53	9.06	4.88	9.56	7.18	2.41	5.15	3.78	3.54	2.63	3.22	2.74	6.42	3.94
35	23.96	25.73	24.92	20.12	20.98	20.55	12.46	11.91	12.10	7.43	12.73	10.16	3.47	5.21	4.28	3.83	3.04	3.55	3.91	6.42	4.73
40	29.22	31.15	29.92	24.60	25.52	24.85	16.14	15.22	15.66	9.98	15.71	12.94	7.14	6.21	6.71	4.46	3.26	4.09	3.91	6.78	4.74
45	33.42	36.16	34.45	28.58	29.74	28.87	20.25	18.28	19.00	12.31	18.07	15.15	8.87	7.18	8.24	6.05	7.60	6.15	4.65	6.78	5.35
50	39.51	41.07	40.12	33.80	34.39	33.96	22.58	22.15	22.60	15.24	20.28	17.75	12.33	9.29	11.12	7.39	9.12	8.04	4.65	6.78	5.64
55	44.36	46.66	45.45	38.57	39.54	39.01	27.51	26.34	26.86	20.04	23.55	21.79	21.76	16.60	19.39	8.70	10.99	9.57	6.09	8.43	6.86
60	51.26	52.01	51.73	44.78	44.66	44.91	32.24	31.52	32.58	24.29	27.48	26.12	23.96	27.43	24.70	10.62	17.66	12.33	8.89	8.60	8.81
65	56.50	56.40	56.69	50.02	49.24	49.86	39.98	37.80	39.13	31.02	31.97	31.79	25.57	27.71	26.60	13.35	20.65	15.35	11.44	18.13	11.87
70	62.22	61.96	62.17	56.95	54.66	55.88	47.42	44.03	45.59	38.66	37.92	38.44	31.93	34.37	32.49	15.66	24.14	18.94	12.03	19.51	14.48
75	67.67	68.51	67.92	62.75	61.51	61.92	52.81	51.87	51.96	45.93	45.58	45.65	36.46	37.94	37.12	17.49	30.63	22.07	12.63	23.10	16.05
80	73.32	75.32	74.22	68.93	68.73	68.73	62.44	57.59	59.95	52.99	55.67	54.10	44.18	42.45	43.21	19.30	41.28	27.50	14.98	27.90	19.19
85	80.01	80.58	80.33	76.47	74.55	75.71	71.51	63.33	67.86	60.87	58.25	59.15	53.56	49.65	51.72	28.26	47.83	33.70	23.55	36.93	24.71
90	87.00	87.52	86.66	84.24	82.50	82.86	79.77	74.33	76.85	73.39	68.02	69.03	59.38	63.59	63.90	37.17	56.32	44.24	29.72	49.81	36.23
95	94.32	93.82	94.09	92.95	91.23	91.91	92.32	88.84	89.81	86.15	84.79	85.45	85.04	88.75	84.73	41.38	82.47	56.49	29.72	76.93	45.66
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

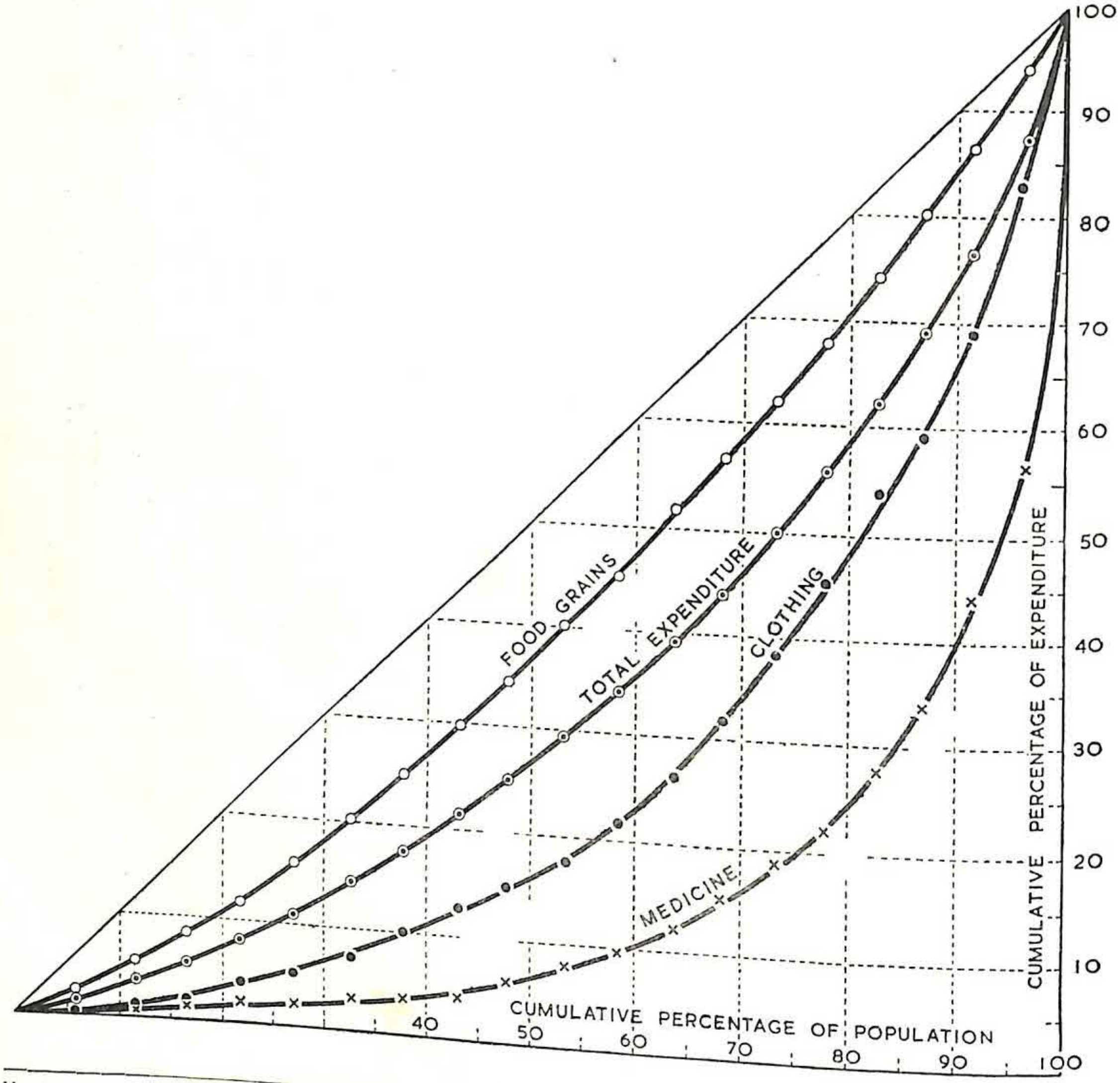
Number of sample villages including 4 uninhabited villages in each sub-sample : 772 (s.s.1), 772 (s.s.2), 1544 (combined).
Number of sample households : 768 (s.s.1), 768 (s.s.2), 1536 (combined).

[1] Excludes Jammu & Kashmir.

CONCENTRATION CURVES FOR CONSUMER EXPENDITURE

Cumulative percentage of total expenditure per 30 days on selected
items based on the National Sample Survey : 9th Round :
May–November 1955 ; All-India Rural

Number of sample villages 1544 including 8 uninhabited villages
Number of sample households 1536



limit of per capita expenditure per 30 days	percentage of household	food grains	food total	total expenditure	sugar	clothing	education	medicine	medical services
(0.1)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			share in percentage of total expenditure						
Rs. 6	bottom 10 p.c.	5.5	4.3	3.5	1.3	1.0	0.8	0.9	0.05
Rs. 13	lower 50 p.c.	40.1	34.0	29.4	22.6	17.8	11.1	8.0	5.6
(Rs. 194)	top 10 p.c.	13.3	17.1	23.2	23.2	31.0	36.1	55.8	63.8

INDUSTRIALIZATION OF UNDERDEVELOPED COUNTRIES — A MEANS TO PEACE*

By P. C. MAHALANOBIS
Indian Statistical Institute, Calcutta

The world population was estimated at about 2,600 or 2,700 million in 1956. About 400 million in the highly industrialized countries like the USA, Canada, Western Europe, and Australia—that is, roughly the NATO powers together with some of the advanced neutral countries of Europe like Sweden, Switzerland, Finland, and Austria—have a high level of living with, of course, a great deal of variation among themselves. These 400 million have a total national income of about $550 (US) billion out of, possibly, a world income of a little over $1,000 billion. The USSR and other socialized countries of Europe with about 300 million people have adopted a planned economy with rapid industrialization. China with 600 million has also started a policy of vigorous economic development. These 900 million in the socialized countries have a total national income, at official rates of exchange, of somewhat less than $300 billion with an average income of about $330 per person per year.[1] The level of living is still low but is rising rapidly and steadily; and there are wide differences from one country to another.

The remaining half of the world's population, or over 1,300 million, are in countries which are generally underdeveloped with a few exceptions, like Japan or Argentina, which have more advanced economies. Their total national income would be about $190 billion which would give something less than $150 per person per year. Nearly 1,550 million in Africa and Asia (excluding Japan and Turkey and a few small places, but including China) have to live on something like $110 billion or on less than $75 per person per year.[2]

The underdeveloped countries often have large resources in the form of minerals and land, which is being or can be used to produce agricultural crops of various kinds, and which offer great possibilities of economic development. The highly industrialized countries in the West already have (and naturally desire to continue to have or to increase) opportunities of exploiting these resources to their own advantage. The "East" (as the socialized countries are called) have large resources of their own and are less dependent on the supply of minerals and raw materials from the underdeveloped countries. However, for both strategic and economic reasons, the Eastern powers would continually try to reduce the influence of the Western powers on the underdeveloped countries, which, so long as they continue to remain underdeveloped, must constitute areas of conflicting interests between East and West.

---

* This paper was presented at the Third Pugwash Conference held in Kitzbuhel-Vienna in September 1958, and appeared in the *Bulletin of the Atomic Scientists*, Vol. XV No. 1, January 1959 and is being reprinted here with the permission of the Editor, Bulletin of the Atomic Scientists—*Editor*]

[1] Figures given in this paper are based on various adjustments and should be used as very rough approximations. Population estimates are subject to many adjustments. National income figures are difficult to compare; and the use of official rates of exchange for conversion into US dollars can be often misleading. The results given here should be, therefore, interpreted as broadly indicative of the general position and should not be used for refined comparisons.

[2] The distribution of national income among countries is extremely uneven. Roughly half of the population (or over, 1,300 million) have $85 billion, that is, just over 8 per cent or one-twelfth of the national income of the whole world. About two-thirds of the population have $160 billion, or one-sixth of the total income. On the other side, the top 5 percent (about 130 million) have $325 billion, or almost one-third of the world income; and the top 10 percent have nearly $450 milliards, or about 43 percent of the total income.

II

The scientific and industrial revolution of the last two centuries was brought about almost entirely by the Western countries, and led to a position of indisputable and undisputed military supremacy of the Western powers. The earlier form of exploitation of the resources of the underdeveloped countries was military and political occupation, that is, by the building up of large colonial empires. In some countries, for example, in China, there was no direct political occupation, but the foreign powers established advantageous economic controls. This indirect form of economic influence, backed invariably by military sanctions, became increasingly more important during the present century so that many underdeveloped countries are politically independent but are economically dominated by an industrialized country.

In this connection I may also refer briefly to the impact of the characteristically Western doctrine of laissez-faire and division of labour at the international level. To put it crudely, there has been a tendency for at least one school of Western economists to assert that it is proper and wise for the underdeveloped countries to specialize in the production of agricultural crops; and to continue to export agricultural products and minerals to be processed by industrially advanced countries which are in a position to utilize these resources with greater efficiency. Experience, however, has shown that it is not possible to improve the level of living beyond a certain limit on the basis of agricultural production alone. The USA supplies a very instructive example. Agriculture is very highly developed, but has to be supported by perpetual subsidies. This being the position in the USA, it would seem practically impossible for any underdeveloped country to attain a high level of living through the export of agricultural products alone. In recent years there has been some recognition of the need of economic development through industrialization of the underdeveloped countries. It has been generally assumed, however, that such economic development would broadly follow the pattern of industrialization in the West, essentially a slow process.

Since the Second World War attempts have also been made through the United Nations Technical Assistance Programme, USA "Point Four", Colombo Plan, etc., to help the underdeveloped countries in economic affairs. The experience of the last twelve years, however, has shown that there has been very little general improvement in the level of living in the underdeveloped countries. It is also being gradually conceded that disparities in the level of living, both relative and *absolute*, between the highly industrialized and the pre-industrial countries have been steadily increasing.

III

The level of living in underdeveloped countries is extremely low, and the distribution of income is extremely concentrated. For example, in India (which is relatively an advanced country in the underdeveloped group), half the population live on less than 10 cents (USA) a day, and only 10 percent of the population have more than 20 cents (USA) per day.

Most of the people depend on agriculture or on indirect income derived from the export of minerals. There are very few modern manufacturing industries. Unemployment, or rather underemployment, is widespread and universal, for lack of capital and modern tools of production. Facilities for education and medical care are extremely meagre. Naturally, there is also a great scarcity of scientific and technical personnel.

A very small group of families or persons have the largest share of wealth, income, and political and economic influence. In fact, the greater the lack of economic development the fewer would be the number of persons who have the effective power of making political and economic decisions. This makes it possible for a foreign power to exert pressure on a small group of powerful persons to give concessions in favour of the foreign power. Such arrangements, because they depend on the will of only a small group of persons, are necessarily subject to violent changes from time to time. Relations between foreign powers and underdeveloped countries are, therefore, basically unstable.

IV

The world situation has changed in a most significant way with the emergence of the USSR and other socialized economies. It is now accepted that economic planning in the USSR (and more recently in other socialized countries) has led to a far more rapid rate of industrialization than had been achieved in West Europe and North America in the past. The Western powers do not any longer have monopoly of ability to supply capital goods and technical know-how. There is also a growing belief among underdeveloped countries that the only way of achieving a rapid rate of industrialization would be through economic planning.

Very recently, the monopolistic superiority of the West in atomic and nuclear weapons and ballistic missiles has also gone. In this situation, it is inevitable that the conflict of interests between East and West in regard to spheres of influence in the underdeveloped countries would become more and more sharp. This, in brief, is the present position. Recent events in the Middle East, for example, corroborate the above analysis.

V

With the progress of industrialization of a country it would automatically follow that its own natural resources would be increasingly exploited by itself. With increasing industrialization it is also inescapable that commercial and economic relations with other countries would increasingly tend to become as between equals.

It is agreed that the very process of economic development would generally broaden the base of social and political decisions. That is, with the progress of economic development a larger and larger number of persons would become involved in making economic and political decisions. External economic and political relations would, therefore, tend to become more stable.

A quick transformation of the underdeveloped countries into industrialized economies would, therefore, reduce the sphere of conflicting interests; and hence decrease the tension between East and West.

VI

The General Assembly of the United Nations passed a Resolution on November 20, 1950 "recognizing that a more rapid economic development of underdeveloped countries is essential for raising the level of productive employment and the living standards of their populations, for the growth of the world economy as a whole and for the maintenance of international peace and security;" and also that there was need of "an increased flow of international public funds" for this purpose. A Committee of Experts was appointed by the United Nations in 1951 "to prepare a report...on unemployment and underemployment in underdeveloped countries, and the national, and international measures to reduce such unemployment and underemployment."

This Committee made a valuable review of the position and reached the conclusion that economic development was the only radical solution of underemployment; and that external assistance, in the form of capital and technical knowledge, was essential for this purpose. The Committee also made a very tentative estimate, on a per capita basis, of capital requirements and came to the conclusion that for about 1,500 million persons in underdeveloped countries (including China) the external capital required would be about $14 billion US dollars (or $9.3 dollars per person) per year and expected that this would lead to a rise of income of 2 percent per capita allowing for growth of population.

In recent years efforts have also been made to give economic and technical aid to the underdeveloped countries through the United Nations Technical Assistance Board, the World Bank, International Monetary Fund, USA "Point Four" (under various names), the Colombo Plan, etc. It is, however, generally agreed that such efforts have not been adequate, and very little improvement has occurred in the level of living of underdeveloped countries. Mention may be made, for example, of Gunnar Myrdal's recent essays and articles in which he has stressed the fact that rich nations are getting richer more and more rapidly, while poor nations are sometimes even getting poorer.

It is also being increasingly appreciated that special efforts have to be made in the initial stage of industrialization. P. M. S. Blackett in his presidential address to the British Association for the Advancement of Science in 1957 stressed the importance of giving external assistance, in the form of capital and technical knowledge, during this initial stage which he called "assisted take-off." He also made some rough estimates of requirements, again on a per capita basis, and thought that about £1 (one pound sterling) per head per year or about 1,000 million or $2.8 billion US dollars per year of external aid would be required for 1,000 million inhabitants for 8 or 10 years for effective industrialization. This estimate of requirement of external aid at the rate of $2.8 US dollars per head per year is less than one-third of the earlier estimate of $9.3 US dollars per head per year made in 1951 by the UN Committee of Experts. This wide divergence shows that these estimates are very tentative; and that much more detailed studies would be required to supply more realistic figures.[3]

VII

Unfortunately, no economic theory of development is at present available to guide our thinking in this matter. It, therefore, seems urgently necessary to start serious and systematic studies to build up a general conceptual framework to handle questions of economic development, and more particularly, to formulate a programme of action (including the supply of external capital and technical aid) to assist the underdeveloped countries.

It is not possible to subsidize the underdeveloped countries indefinitely. The aim must be to make each country sufficiently developed to reach some kind of a steady stage or a balanced economy in the course of say 10 years (or 15 years at the most) so that the inflow of external aid would gradually decrease and, preferably, cease after 5 or 10 years. That is, it is necessary to make some realistic estimate of the total external aid required for this purpose so that the industrially advanced countries can get some idea of the magnitude of the task.

---

[3] It is worth noting that the expenditure on defense is of the order of $100 billion (of which the share of NATO powers is about $60 billion and of the USSR about $25 billion) per year at present; if expenditure on associated research is included the total would be higher. It would be seen, therefore, that the requirement of about $3 billion (as estimated by Blackett) or about $10 billion dollars per year for 1,000 million (on the basis of the rate used by the UN Committee) would amount to no more than 3 percent or 10 percent of the defense budget of the world. A saving of 3 percent (or 10 percent, as the case may be) in the cost of defense, if used in a wise manner, would be more than sufficient for the economic development of 1,000 million people in the most backward countries of the world.

Estimates prepared by the UN Committee of Experts in 1951 and by P. M. S. Blackett in 1957 were both on a per capita basis. Some broad general considerations would, however, show that capital requirements per head would differ from one country to another. The bigger a country (as measured by its geographical area, unexploited natural resources, and population) the greater would be the possibility of achieving its industrial development out of its own resources. In a big country it is possible and desirable to push back the manufacturing process to the utmost limit in order to expand continually its capacity to make investments increasingly out of its own domestic resources.

We may consider the case of India as an example. In India it would be economical to establish a heavy machine building industry which would manufacture heavy machines and equipment required for the installation of factories for the production of steel, fertilizers, aluminium, etc., or for the production of heavy electrical equipment like big generators, transformers, switchgear, etc. It would be also economical gradually to establish large-scale industries for the manufacture of synthetic raw materials of many kinds (including the production of petrol from coal in case an adequate supply of oil is not discovered). Once such basic industries are established it would be possible to expand the production of electricity, coal, steel, aluminium, fertilizers, mining and transport equipment, etc., and then, with the help of such heavy machinery, producer goods, and energy to manufacture machinery for the increasing production of consumer goods. It follows that a big country would require a comparatively small amount of outside capital.

It is of interest to note that in the case of China, with a population of roughly 600 million, the Soviet loan of roughly $1.25 billion US dollars, given for economic development during the first Five Year Plan (from 1953 to 1957), represented only about two US dollars per head spread over five years, or forty US cents per head per year. This seems to have been quite enough to make China "take off" on its way to rapid industrialization, but is clearly a lower limit of external capital requirement.

For a big country, external economic transactions (of which foreign trade is a most important form) would be comparatively small in relation to the magnitude of its internal economy. In the USA, for example, external trade is only about 5 or 6 percent and in the USSR only about 2.5 percent of the national income. It should not be very difficult, therefore, for a big country to attain a fairly stable balance of payment.

It is clear that planning must have a fairly long perspective of time. Factories can be established for the production of consumer goods in 2 or 3 years; large-scale production of steel and other metals, electricity, fertilizers etc., would take at least 7 to 10 years; heavy machinery building and heavy electricals would take from 10 to 15 years. The most slowly maturing sector would be that of scientific and technical manpower which would require planning for 25 to 30 years.

IX

In the case of a small country (as measured by its area, natural resources, or population) the problem is more difficult. There is a limit beyond which the manufacturing processes cannot be pushed back, because there is a minimum factory size below which production is uneconomic, in the case of many important commodities like heavy machines, steel etc. It may not be economical for a small country to establish a heavy machine manufacturing industry, because it would not be possible for the country itself to utilize the output every year. There is also a limit below which it may be difficult to undertake in small countries the economic production of steel and other metals on a large scale. A small country, therefore, would have to depend, to a larger measure than a big country, on the import of capital goods as well as of many producer or consumer goods. The industrialization of a small country (unless it has large natural resources like oil or other minerals which it can export on profitable terms) would presumably require comparatively more external aid.

Also, the smaller a country the greater is likely to be the magnitude of its foreign trade in relation to its whole economy. In the case of a small country it is, therefore, of great importance to consider the expansion of foreign trade in desirable directions. In a paper prepared for the Bandung Conference of Afro-Asian countries in April 1955 I had given some preliminary consideration to this problem (see Appendix).

The pattern of development is likely to be more or less similar for big countries. But it is not possible to reduce the same pattern mechanically to scale to suit the needs of small countries. The developmental plan for a small country must have specific relation to its own natural resources and also to possibilities of expanding its foreign trade in cooperation with other countries of the same region.

X

It is suggested that attempts should be made to set up, as soon as possible, small groups of experts to study the problem of industrialization, at a concrete level, for the underdeveloped countries. The aim would be to formulate the broad strategy of planning for economic development over a period of 15 or 20 or 25 years. This would imply preparing a rough time programme for the development of agriculture and the establishment of modern industries with appropriate priorities. Special attention will have to be given to health, education, and the supply of scientific and technical manpower.

Consideration will have to be given to the expansion of foreign trade, not merely as an extrapolation of current trends, but with a changing pattern over time which would be in keeping with and would also promote rapid economic development. Naturally, economic relations with other countries, especially the underdeveloped countries in the same region, will have to be considered; and economic measures will have to be devised which would promote both the expansion of trade and the economic development of the countries concerned in the regional group. Gunnar Myrdal, for example, in a recent lecture in India (1958) has pointed out the advantage of a group of underdeveloped countries in a particular region forming, by mutual agreement, a "common market" protected from outside with, however, a free flow of trade and specialization of manufactures within the protected area.

Studies will also have to be made of not only the amount but also of the pattern of external aid, in the form of capital and technical knowledge, which would have to be supplied by the more highly industrialized countries to start the process of industrialization and to carry it through until some kind of a balanced economy has been established when no further special subsidies or long-term developmental loans would be required. Formulation of even rough estimates would be of great value in supplying a starting point for further thinking.

It is believed that enough information is available for at least a good number of countries to enable such studies to be initiated immediately. It would be necessary, of course, for experts from the advanced countries to supply the leadership. At the same time, it would be essential to associate with these studies, as early as possible, technical personnel from the underdeveloped countries. In fact, one important aim of such studies would be to provide opportunities to personnel from underdeveloped countries to acquire knowledge and experience of planning for economic development.

The task proposed in this section would call for patient study and the collection of much essential information. It may be necessary to carry out special surveys for this purpose. The formulation of a programme of surveys with indication of priorities would indeed be a most useful piece of work as a first step for development. The approach will have to be of a "spiral" type. Attempts would be made to make rough estimates on the basis of available data. This very process would indicate gaps in information. As these gaps are filled up, the first estimates would be revised; and the process of revision would disclose the need of further information; and so on. It is, therefore, necessary to begin serious studies as early as possible.

XI

It is true that mere formulation of even a sound technical plan for development would not be of any use until social and institutional conditions within a country become favourable to industrialization. But the very formulation of a general conceptual framework for economic development would serve as a powerful stimulus to set social forces in motion for industrialization.

A number of countries in Afro-Asia, such as India, Pakistan, Burma, Ceylon, Indonesia, Egypt and Syria (United Arab Republic), etc, have already started planning for economic development and would welcome and greatly profit by the studies of the type discussed in this note. In fact, these countries would probably agree to help and gladly participate in such studies. There are many other countries which would be seriously interested.

There are also countries in which there is as yet no definite movement for economic development. But, in such countries also, individuals here and there may have already started thinking or may soon begin to think on this subject. The proposed studies would give them encouragement and guidance and thus help in the creation of favourable conditions for industrialization.

XII

With the progress of industrialization, disparities, both absolute and relative, between the highly industrialized countries and those which are at present underdeveloped, would decrease. Also, some of the advantages which the highly industrialized colonial powers (in either the political or the economic sense) now enjoy in having preferential access to sources of cheap raw materials or of markets would, no doubt, decrease; but this does not mean that there would be any worsening of their absolute level of living. On the contrary, with fuller and better exploitation of the resources of all the countries of the world, it should be possible to reconcile conflicting interests in such a way as to safeguard the level of living of all the countries of the world.

The Western Powers have so far refrained from promoting industrialization of the underdeveloped countries in any significant way, possibly because they do not wish any competition to grow about securing minerals and raw materials at concession rates or to lose markets for manufactured products. It is true that Western Powers, especially the USA. have been giving aid to the value of billions of dollars every year to some of the underdeveloped countries, but such aid has been extremely selective and mostly for military purposes. Although there is much talk of economic aid, the actual amount devoted to industrial development is extremely small and forms an insignificant proportion of the total aid. For example, the total budget for economic and technical assistance through the United Nations and the Specialized Agencies like the International Labour Organization (ILO), Food and Agriculture Organization (FAO), World Health Organization (WHO), etc. amounts to only about $34 million dollars compared to something like one hundred times that amount spent on bilateral aid mostly for military purposes. The primary object of Western aid being of a political or military nature, stringent conditions are usually attached.

The USSR, on the other hand, may be eager to bring about the industrial revolution in the underdeveloped countries as quickly as possible, because this would make them independent of the Western Powers in economic affairs. This would also create a working class which, according to the Marxist view may promote social revolution from within. As the USSR and other Eastern Powers do not have any direct military or political objects in view, they can and actually do give economic aid without strings. Although it is only about two or three years since the USSR has started giving economic aid in any large measure, the impact on the underdeveloped countries has been already very great.

The basic difference in the outlook of the West and the East has important consequences. Western aid, primarily military in nature, is usually welcomed by being given to small groups of persons who happen to be in power, often against popular will; and thus serves to preserve reactionary social and political conditions and hamper industrialization in countries receiving such aid. The aid from the USSR and China on the other hand, is being increasingly sought by and given to groups who are more eager for economic progress and, therefore tend to have greater popular support behind them.

However once it is realized that the industrialization of the underdeveloped countries is inescapable, and also that any country which actively helps in the process of transformation would steadily gain in political and economic influence it is inevitable that *both* blocks would try to assist in the process. The advantage will lie with that block which approaches this task with greater honesty, sincerity and efficiency.

Nuclear weapons and guided missiles have made the cold war obsolete. Stopping of nuclear tests and gradual disarmament would be most desirable and welcome, but would not by themselves guarantee peace. In the future peace can be preserved only by active cooperation between the two blocks (and this may be possible in principle, even without complete or drastic disarmament. Such cooperation would be difficult if rivalries continue to persist in the underdeveloped countries. A quick transformation of the underdeveloped countries into modern industrialized economies would eliminate areas of conflicting interests.

In any case, the proposed studies would give guidance to industrialized countries in the efficient utilization of their economic and technical aid to underdeveloped countries. With the progress of industrialization, rolling adjustments would tend to be made in areas of conflicting interests. Gradually, it may be hoped, both blocks would begin to appreciate the advantages of cooperative efforts for the industrialization of underdeveloped countries as an indispensable condition for world stability and peace.

Appendix

ECONOMIC DEVELOPMENT OF AFRO-ASIAN COUNTRIES

(Prepared for Bandung Conference, April 1955)

The enduring basis of Afro-Asian understanding must be ultimately established on mutual cooperation to bring about a steady increase in the level of living through economic, social, and cultural development in the Asian and African countries. Anti-colonial and peace movements must be looked upon essentially as attempts to secure favourable conditions for such developments.

The aim must be to promote the exploitation of all available natural and human resources for the national development of each individual country. The Afro-Asian region or even Southeast Asia as a whole has such large resources that there is not the slightest difficulty, at the technological and economic level, to attain a rapid rate of growth of all the national economies. Much capital goods and technical aid from the more advanced countries would be required in the beginning. Fortunately, with the improvement of international relations, such aid is likely to become increasingly available from different parts of the world. It would be of advantage if such aid could be channeled through the United Nations or similar international agencies.

It is, however, not necessary to wait for UN to take action. It is possible and, indeed, desirable to initiate action at the level of the countries concerned. For example, a Standing Technical Committee (with a small permanent secretariat) may be appointed to study possibilities of fostering economic, social, and cultural cooperation among the Bandung countries. The work of the proposed Committee would be mainly concerned with the study of problems of long-range development. This would not duplicate but would be complementary to activities which are directed to solving more urgent problems in the Economic Commission for Asia and the Far East (ECAFE) and the Colombo Plan.

The proposed committee might, for example, examine possibilities of long-term industrial developments in the countries concerned, and possibilities of expansion of international trade not on the basis primarily of the current pattern of imports and exports but from the point of view of promoting a changing pattern of international trade deliberately directed to the national development of all the countries concerned. India at present has a good export market for cloth and some other manufactured products. In a static approach, the current pattern of trade would be sought to be maintained which would retard industrial development in the importing countries (as had happened in India in relation to the United Kingdom). In a dynamic approach, India would actively promote a rapid industrialization of the more backward countries, fully recognizing that the pattern of Indian exports must change thereby but also appreciating the possibilities of a steady expansion of the foreign trade of all the countries concerned at increasing levels of manufactures.

The dynamic approach necessarily calls for a wide horizon of time. Trade agreements and understandings have to be visualized as extending over a period of 5 or 10 years or even more. Prices, terms, and conditions must have much greater flexibility than short period trade contracts. The aim would be to maintain at an agreed level or expand in a suitable way, over a number of years, the total volume of imports and exports (measured, say, at constant real prices); and, at the same time, to try to change the pattern of trade in such a way as to promote increasing industrialization of the countries concerned. For example, if India can be assured of a steady supply of, say, rice from Burma over a number of years (preferably at constant real price), then India may undertake to supply not only manufactured consumer goods but also to help in the industrial development of Burma through the supply of minerals, raw materials, capital goods, and technical assistance. In this approach India must give up the idea of earning large profits through the continuing export of particular types of products; and adopt the policy of mutual help and benefit to attain increasing levels of industrial development of both Burma and India.

A dynamic policy of mutual development over a number of years can be promoted most effectively at government level with, however, the fullest participation of the private sector which would be entirely in keeping with the economic policy of India.

As already mentioned, a first step may be to set up a small working group or technical committee to initiate economic, social, and cultural studies to supply a scientific basis for preparing long-term developmental schemes on a regional scale. A large membership is not essential; and work can be started with a small group of countries or even on a bilateral basis by agreement between two countries.

To make a beginning, India may offer certain physical facilities like accommodations, library, and a small professional and office staff to serve as the nucleus of a secretariat. It would be desirable to have a director from outside India; and the professional staff must come chiefly from the participating countries.

FOOD RESOURCES AND POPULATION OF INDIA
A HISTORICAL STUDY

By P. C. BANSIL
Planning Commission, India

SUMMARY. A historical study of the production of food and population of India has been presented in this paper broadly under five periods, viz, the separation of Burma, the second world war, the partition of India, the decontrol and the present position. Figures of production, export and import of food in the different periods have been studied and reasons of increase and decrease in exports and imports and shortages and rising prices of food in the respective periods have been analysed. An attempt has been made to establish that the shortages and rising prices are not due to pressure of population but to many other factors. The supply of foodgrains today is not the same thing as total production and demand is not the same thing as consumption. Marketable surplus in the restricted sense has suffered. The effective demand that of the nonproducers has at the same time tremendously increased. This imbalance has created the existing disequilibrium in the forces of effective supply and effective demand resulting in rising prices and temporary shortages. The food potential of the country is otherwise pretty high and there is nothing to be pessimistic.

1. INTRODUCTION

It has generally been accepted that the production of food in India was keeping pace with the increase in population upto the end of the thirties. The Census Report, 1951, however, calls the year 1921 as the 'Great Divide' because India was a net exporter of foodgrains before this date but became a net importer after 1920. There being apparently no controversy about the food and population relation for the period upto 1920, we will restrict our present study to the period from 1921 onwards.

In this paper we propose to explore whether a study of India's emergence as a net importer of rice and the drying up of her wheat exports from 1921 could give a sufficient proof of the pressure of population on her food resources.

2. THE PERIOD BEFORE THE SEPARATION OF BURMA

Rice imports. Before separation, Burma was the main supplier of rice to India but rice imports from Burma did not follow a steady pace. Imports fell down from one million tons in 1921-22 to nearly half a million tons both in 1923-24 and 1924-25 and reached the peak figure of 2.2 million tons in 1934-35, and again came down to 1.7 and 1.6 million tons in the following years.[1] This downward trend in imports continued in the first two years after separation. One of the many factors responsible for these fluctuations in rice imports was that the Burma rice was of a relatively low quality and the cheapest variety. It catered to the taste of poorer section of South Indians whose favourite dish was 'Kchichri' (a preparation of rice and pulses).

Again the study of the price differential of low grade rice in India and Burma, according to the report on the marketing of rice,[1] will show that Burma rice was invariably cheaper as compared with the coarse varieties of Madras and Bombay. While the price per maund of such coarse varieties in Burma varied between Rs. 2-11-7 and Rs. 1-3-5 in the year

---

[1] Report on the Marketing of Rice in India and Burma, 1941, pp. 564-67.

1934-35.[1] The average price of a coarse variety like 'Kusuma' in Madras and 'Pattani' in Bombay was Rs. 3-5-4 and Rs. 3-15-2 per maund respectively in 1934-35. The Burma rice was being sold at Rs. 3-5-6 per maund, in the city of Madras, whereas local rice would cost Rs. 4-5- per maund. As for the imports from other countries they also consisted mainly of broken rice of the cheapest variety.

Since the supplies of cheap rice were available from outside, the Crop Planning Sub-Committee, 1934, recommended that no more area should be brought under rice.[2] The various provinces, therefore, made every effort to encourage cash crops.[3] But the area under such crops improved only slightly, while that under rice remained practically stationary. In the absence of any alternative demand for cash crops, the cultivator found it very difficult to reduce cultivation of rice when the prices were falling particularly when there was no appreciable demand for any competitive crop.

World demand. From 1909-13 to 1936-37, average world production of rice (excluding China) rose from 50.4 to 64.4 million tons, i.e., by about 28 percent.[4] The countries mainly responsible for the increases were India, Burma, Siam, Japan and Korea. Rice is consumed mainly by the countries of Asia. In Europe, France is the chief importer of rice, that too from Indo-China. The World market in rice being thus tight, the Government of India made efforts to find market for Burma rice in India. So much so that a protective duty of twelve annas per maund was imposed on all foreign imports of rice with effect from April 1, 1935.[4]

This had the desired effect of shutting off to a great extent the imports of broken rice, chiefly from Thailand, in which case they steadily declined from 283,000 tons in 1934-35 to 18,000 tons in 1935-36.[5]

The real position. Rice imports as we have already seen did not keep any steady pace. Area under cultivation of rice varied only from 80.3 to 84.9 million acres for India including Burma and 70 to 72 million acres for India alone during the period 1920-21 to 1936-37.[6] This shows that though India depended on Burma for the supply of cheap rice, it would be wrong to say that she could not produce at home her normal requirements or that the increasing pressure of population was in any way responsible for these imports.

The real position on the other hand appears to be that Burma had a large exportable surplus of rice for which there was no world demand. Cost of production in Burma was low because of cheap labour. The land there was also not fit for the cultivation of any other crop. When it formed part of India there were no trade restrictions between the two areas. As most of the Burma business was in Indian hands, they found it convenient to trade with India. Consequently, half of the Burma trade, both import and export, was with India. It would thus be no wonder if it were presumed that India was rather obliged to get her rice supplies from Burma to adjust the economy of the country as a whole. It would otherwise be seen that though importing rice from Burma, India from 1920-21 to 1936-37 was a net exporter of foodgrains. Interestingly this is so even when the survey is restricted to exports from India excluding Burma. Rice imports from Burma in that case are to be treated as internal trade.

---

[1] Report on the Marketing of Rice in India and Burma, 1941, pp. 564-67.

[2] Baljit Singh : *Population and Food Planning*, p. 97, (Bombay, 1947).

[3] Brij Narain : *India Before and Since the Crisis*, p. 385 Vol. II, (Allahabad, 1939).

[4] Rice Report : *Op. Cit.* p. 2.

[5] G. B. Jathar and S. G. Beri : *Indian Economics*, Vol. I, p. 137, (1949).

[6] Rice Report : *Op. Cit.*, p. 100.

3. WHEAT POSITION

Just as India's food imports consisted mainly of rice, her chief exports were of wheat, although she was also importing wheat mainly from Australia.

They showed wide fluctuations in the imports and exports of wheat from year to year. Wheat imports were maximum for the years 1921-22 and 1928-29, while exports after touching the million mark in 1924-25 continued to fall. This fall is said to be the result of increasing demand for wheat at home. But abnormally heavy imports were said to be the result of poor crops in the preceding years.[1] Following the same logic, however, it remains unexplained how the bumper crop of 1929-30 failed to slash down imports to near about the pre-1928 level. Imports or exports of wheat would not, therefore, appear to bear any relation to production at home. Let us then examine the factors which governed the imports and exports of wheat during the period.

Imports of wheat. A study of wheat market during the 'thirties' and 'forties' shows that the home market was much depressed and the position with regard to foreign imports became so alarming that the Government was forced to impose a special duty of Rs. 2 per cwt. on wheat and Rs. 2-8- per cwt. on wheat flour with effect from April, 1931 in the interests of growers.[2] When even this failed to improve the situation, imports of wheat were banned under the Wheat Act of 1931. This, too, could not check the downward movement of prices.

The price per candy (656 lb) of Australian wheat in the United Kingdom (C.I.F. Liverpool and / or London) in the years 1933-34, 1934-35 and 1935-36 was Rs. 19, Rs. 20 and Rs. 20.6 respectively. The corresponding quotations at Karachi (ex-store) were Rs. 23.6, Rs. 21.8 and Rs. 23.8 respectively.[2] This difference in prices becomes greater when it is realised that it was cheaper to ship wheat from Australia to Calcutta or Bombay than to transport it all the way from the Punjab or U.P.[3] According to the Central Banking Enquiry Committee, freight per maund of wheat from Australia to Calcutta was annas 6, while the railway freight from Lyallpur to Calcutta was Rs. 1-3-3.[4] This explains why Australian imports continued in spite of the import duty levied by India and probably there being an exportable surplus in India itself. Wheat imports would thus seem to be directly related to its price parity in India and Australia rather than any effective home demand.

Export of wheat. India's export trade in wheat with the Western countries dates back from 1870. Starting from 3.9 thousand tons in 1869-70 it touched the million mark within a short period of about 14 years. They, however, showed wide fluctuations from year to year, resulting in their drying up by 1935-36.

---

[1] Report on the Marketing of Wheat : *Op. Cit.* p. 40.

[2] Adapted from G. I. Kovoor : *Rural Reconstruction Through Production and Marketing of Wheat in India*, 1924, p. 280.

[3] Dr. A. I. Qureshi : *The Present Food Situation in India*, Hyderabad, 1944. pp. 8-9.

[4] *Op. Cit.* : 1934, p. 68.

Purchasing power. It is, however, said that these exports did not represent any real surplus but owed their existence to the poverty of the people who could not afford to purchase their normal food requirements.[1] This contention seems plausible on the face of it and most of the early writers were led away by it.

A deep study of the question, however, proves this conclusion wrong. Even otherwise opinions have been expressed by experts like Bowley and Robertson that "In India, as in the rest of the world, there has apparently been in recent years an 'overproduction' of agricultural as compared with manufactured goods.... The food-producing village, and even the food producing family is a self contained unit; and in some cases, whatever its privations in other respects, its consumption of food has probably increased in consequence of the fall of prices."[2]

That these exports were not merely due to the lack of purchasing power of the cultivator has also been shown by Voelcker. He reached the conclusion that such exports were real in the sense that this extra production would not have been there, if the foreign demand was removed.[3] Entirely different varieties were grown according to him for the home consumption and the foreign market.

The price differential. The Report on the Marketing of Wheat has given the explanation that our exports increased in those years when the price difference between wheat and barley or wheat and gram was the greatest so that the poorer people could release more of wheat for export. This can only be regarded as one of the contributing factors. All the same more cogent reasons have to be sought for this unsteady movement of exports.

According to the Report on the Marketing of wheat though the price of wheat in India was the lowest, yet it could not compete with the larger exporting countries of the world. "In recent years price considerations have prevented export on any appreciable scale for in spite of a preference, under the terms of the Ottawa Agreement, of 2 shillings per quarter (480 lb) Indian wheat has very rarely been cheap enough with Australian, Canadian and Argentine wheats in the United Kingdom markets."[4] The Wheat Conference held at Simla in May,[4] 1934, also pointed out that "there is no export because for some years our prices have been above world parity."

During 1933, for example, the price quotation of Indian wheat in London was Rs. 1-14- per maund at Karachi Railway Station—a price far below the lowest figure ever reached during the depression. Against this low quotation, the price of wheat in the harvest season of that year at Lyallpur, was Rs. 2-2- per maund.[5] Even if all incidental expenses are ignored, the Punjab wheat at Karachi could not be delivered at less than Rs. 2-13-8 per maund as the Railway freight from Lyallpur to Karachi was Re. -/11/8 per maund.

---

[1] *C. f.* Dr. Gyan Chand : *India's Teeming Millions*, 1939, p. 298; Ranadive : *Population Problem of India*, 1930-p. 126; Chowdhry Mukhtar Singh : *Rural India—Peasants Poverty, its Causes and Cure*, Allahabad, p. (xiii).

[2] *Report on the Improvement of Indian Agriculture* : John Augustus Voelcker. 893, pp. 294-95. [The same is the opinion of Albert Howard and Gabrialle L.C. Howard : *Wheat in India*-1909. p. 45].

[3] Report : *Op. Cit.* p. 44.

[4] Proceedings, p. 1. Also Supplement to the Report on the Marketing of Wheat in India, 1946, p. 9; and Jathar & Beri, *Op. Cit.* p. 140.

[5] G. I. Kovoor : *Op. Cit.* p. 39.

Naturally the question of exporting any wheat to London, the chief importer of Indian wheat did not arise under the circumstances. The result was that wheat exports during 1932-33 and 1933-34 slashed down to the lowest figure of 2 thousand tons annually.

Price parity, it seems, was playing an important role even in the case of internal trade. Calcutta market had been the natural and practically the only outlet for the export of wheat from the United Provinces. In the years of slack European demand for Indian wheat, the market was captured by the Punjab to the detriment of U.P. With the coming of Australian wheat even the Punjab was ousted from the Calcutta market.[1]

This would mean that the responsibility for the fall in the Indian wheat exports lay more than anything else in her inability to compete in the world wheat market. If some wheat was still being exported against all these odds, it was due to the quality[2] of Indian wheats and the time when it was available in the European markets.

Wheat Conference. In the presence of such a world situation, India was faced during the period with the problem of encouraging more consumption of wheat at home. While speaking at the Wheat Conference of 1934 Jenkins pointed out that one of the reasons militating against an increased consumption of wheat in India was the large quantity of cheap rice which was being imported from Burma.[3] In the concluding remarks, he again emphasised that the only solution of the problem lay in fostering home consumption and that could be done by a general and indiscriminatory reduction in freight rates, and by starting a publicity campaign on 'Eat more wheat' lines.

The Hon'ble D. J. Boyde could see only two possible ways out. One was the increase in the internal demand for wheat, and the other was of prohibiting any increase in area under wheat.[4] The Chairman suggested that the Punjab should consider the possibility of diverting 5 percent of its irrigated area under wheat to oil-seeds.[5] The problem was practically of all India importance. Representative of the Central Province confirmed that there was a good deal of wheat in his province and that production and consumption balanced with a 10½ anna crop and an acreage of 3,400,00.[6] He urged that nothing should be done which may result in the Central Province's wheat being displaced in the local markets by the Punjab or Sind Wheat.[7]

Again, to achieve the same end, the Crop Planning Conference in 1934-35 decided that no attempt should be made to produce more than 9½ million tons of wheat in the country and that there should be no stimulus to increase the area under wheat. The future of the crop thus depended on world supplies and on India's policy regarding wheat.[8]

Import of foodgrains and its causes. Our analysis in the preceding pages is sufficient to refute the accepted theory that India's emergence as a net importer of foodgrains was merely due to the pressure of population. As we have discussed at length, imports or exports of foodgrains were governed by factors other than production or consumption in India.

---

[1] Bulletin No. 4 of the Bureau of Statistics and Economic Research, United Provinces, quoted in the United Provinces Wheat in 1938-39 and *Days to Come*, J. K. Pande, 1939, p. 7.

[2] Chowdhry Mukhtar Singh : *Agrarian Reliefs*, p. 187 and Kovoor : *Op. Cit.* pp. 19-20.

[3] Proceedings of the Report : *Op. Cit.* p. 5.

[4] Report : *Op. Cit.* p. 12.

[5] *Ibid.* p. 16.

[6] *Ibid.* p. 7.

[7] *Ibid.* p. 13.

[8] Quoted by W. Burns : *Technological Possibilities of Agricultural Development in India*, 1944, p. 53.

If population was increasing, production was also increasing which cannot, of course, be established on the basis of existing data. We find that irrigated area during the first thirty years of the century had increased by nearly 30 percent. Various land improvement schemes launched by the Government must have also added to the existing food resources to some extent. But the general depression of the 'thirties' placed Indian agriculture in doldrums. It being uneconomical for the Indian peasant to grow more food just for the sake of dumping it in his 'kacha' store house, he lost all incentive to put his heart and soul in agriculture. If the marginal food requirements of certain parts of the country were met by imports, it does not mean that they were dictated by the existence of a 'Food problem' as such within the country. The position on the other hand was that home potential could not be exploited fully because of the lack of an effective demand.

4. THE PERIOD AFTER THE SEPARATION OF BURMA

Burma was separated from India in 1937. In the first two years after separation, rice production in India fell to 28.8 and 26.1 million tons respectively from 30 million tons in 1936-37 and from an average of 28.0 million tons during the previous quinquennium.[1] In Burma, however, production of rice rose to 5.0 million tons during 1937-38 as compared with 4.85 million tons in the quinquennium 1932-33/1936-37.[2] Though the production of rice fell in India against a slight rise in Burma, rice imports from Burma also fell to 1.3 and 1.4 million tons in the years 1937-38 and 1938-39 respectively, against the average of 1.7 million tons during the previous quinquennium. Again, these imports fell when India figured prominently in the import and the export trade of Burma (both accounting for more than 50 percent of the share[3]) and there was free trade between the two countries.[4]

Although two years is too short a period to enable us to come to any conclusion about this fall in imports, it can presumably be ascribed to the Indian market being not responsive to the imports of rice from Burma. Had it not been for their cheapness,[5] rice imports from there might have totally dried up. These cheap imports made the position so serious that even after the Second World War had broken out a meeting of the Rice Committee of the Imperial (now Indian) Council of Agricultural Research held in New Delhi on November 28, 1939 passed a resolution to the effect that "suitable steps should be taken to maintain the price at a reasonable level and for this purpose, the Government of India should be moved to consider the advisability of regulating imports of rice."[6]

As a result of this, officers of the Central Marketing Staff held a series of informal discussions with the various trade interests concerned at Calcutta, Rangoon, and Madras in December, 1939 and January, 1940. It was found that Calcutta and Madras being important focal points of the rice trade, vitally affected the whole rice position and are, in turn, affected by imports from Burma. From the informal discussions which took place at the time, it was accepted by the rice millers at Calcutta and Madras that the interests of Indian paddy growers would be safeguarded if the price of Burma rice imported in to India was maintained at a reasonable level, viz., Rs. 3-8- per maund for Small Mills Specials ex-dock Calcutta and for broken rice of a quality equivalent to Broken No. 3 delivered at Madras. This was thought to be feasible if Burma undertook to control exports to India by means of licences and reduced or increased shipments to India according as the price of Small Mills Specials fell below or rose above the parity of Rs. 3-8- at Calcutta. Millers and traders in Rangoon, were also not opposed to the idea of regulating imports of rice provided steps were taken to give Burma a larger share of the Indian market by imposing fairly heavy duties on foreign paddy and rice.[1]

---

[1] *The Food Statistics of India* : Government of India, 1946, p. 29.

[2] Estimates of Area and Yield of Principal Crops in India, 1938.

[3] Foreign Commerce Year Book, 1939, based on Seaborne trade of India.

[4] J. Russel Andrus : *Burmese Economic Life*, American Institute of Pacific Relations, 1947, p. 167, for various other favourable circumstances.

[5] Rice Report : *Op. Cit.* pp. 564-65.

[6] Rice Report : *Op. Cit.* p. 198.

We do not know what the course of events would have been if the War had not taken a serious turn. It is, however, clear that imports even during this period were dictated by the pricing factor rather than by anything else.

Imports and exports of wheat also continued to be governed by the price factor as before. There was a small recovery in the international wheat trade in 1936-37 and owing to favourable price parity India too got her share. Her exports went up to 285 thousand tons in that year and 539 thousand tons in the following year, a figure which had never been reached during the last decade. The chief reason for these heavy exports might be the falling wheat prices in India after March, 1937 when the duty on imported wheat was not renewed. It was reimposed only on December 7, 1938 when the internal prices showed a sudden rise. The presence of a bumper wheat crop of 1938 in Canada and America helped further to oust Indian wheat from the foreign market.[2]

That the imports and exports of Indian wheat in 1938-39 depended upon its price will be clear from a study of the Bureau of Economics and Intelligence, U.P. during this period.

It shows that the price margin for Karachi wheat in London began to close up after August, 1938, so much so that it remained only a few annas in September, 1938 and ceased to be quoted after October, 1938. It is also quite clear from the same study that India's wheat exports followed these fluctuations in prices.

When wheat exports dwindled from August, 1938, there was a corresponding increase in its imports. This was because of the inability of Indian wheat to compete successfully with the Australian wheat which not only drove out Indian wheat from European markets but also offered competition even in the Indian market, with the result that the imports of wheat into India went up from 34 thousand tons in 1937-38 to 165 thousand tons in 1938-39.[3]

This analysis of India's foreign trade in wheat and rice would show that it was always governed by the pricing parity prevailing in the international grain market.

---

[1] Rice Report : *Op. Cit.* pp. 198-201.

[2] Jathar and Beri : *Op. Cit.* p. 140.

[3] J. K. Pande : *United Provinces Wheat, Op. Cit*, p. 6.

5. The second world war and food

Early in 1939 War clouds were hovering over Europe and every country was keen to stockpile foodgrains. This will perhaps explain the rise in Indian imports of foodgrains during the year, 1939. Rice imports alone amounted to 2 million tons, the highest in the import history of India before 1939.

Immediately thereafter rice imports fell below the pre-war average. They came down to 1.3 and 1 million tons in the following two years. War had not spread in the Eastern theatre as yet and there was practically no difficulty in importing foodgrains from South East Asia, particularly Burma. But this fall in rice imports seems paradoxical when we study the production figure of rice in India and other SEA countries. Rice production in India in 1940-41 stood at 23.8 million tons, the lowest on record, during the previous decade. In Burma and Siam, however, production was 5.7 and 3.4 million tons[1] respectively, the highest ever recorded there. This would seem all the more paradoxical when it is realised that pressure on internal resources had also increased due to the demand of the army.

There were also other war-time demands on India's food resources. She had to support foreign troops stationed on Indian soil and to meet certain demands of rice and wheat from her neighbouring countries like Ceylon and her own troops serving in Iran, Iraq and the Middle East.

Moreover, the trend towards urbanization and employment opportunities provided by the war attracted people to the cities. Speculative hoarding of large stocks by consumers, producers and merchants also aggravated the situation. Coupled with this was the tremendous wastage of food. Although there was no actual fighting on the Indian soil, the bombardment of the eastern borders resulted in the destruction of large stocks of food.

Rationing introduced in 1943 also resulted in wastage of food. Though year to year figures of the wastage are not available an idea can be formed from an estimate of the Food Grains Investigation Committee 1950 that as much as 1½ lakh tons of foodgrains were lost in transit in 1949.

Food supplies. Against all this accentuation of demand, the supplies of food from sources outside India were perhaps the minimum. India was actually a net exporter of rice during the years 1942-43 to 1945-46. Again, there was a sort of a permanent shift in the food trade in the sense that India became a net importer of wheat instead of an exporter from the year 1943-44. These imports though not very large during the period under review were steadily increasing. They exceeded even those of rice in the post-partition years. With a record wheat crop of 11.2 million tons in 1942-43—a year when her net exports of foodgrains were of the order of 3 lakh tons—the emergence of India as a net importer of wheat from the very next year—1943-44 is a matter which needs investigation.

Before we look into the causes for this strange phenomenon it would be interesting to note that before the War, for whatever reasons, India had been getting food supplies of the order of 1 to 2 million tons from sources other than her own. If we can prove that she could survive this period of trial without any untoward happening in the country, we would be justified in asserting that food imports during the earlier period when the number of people required to be fed were surely less, were not required to meet any inherent demand of the people. The Bengal Famine of 1943 may be quoted as the biggest calamity of the period. The blame for the famine, however, lay not in any real food scarcity but elsewhere. In fact the position in the country from the point of view of food stocks was thought to be quite normal.

---

[1] Rice Statistics, February, 1948, Directorate of Economics and Statistics, Government of India, p. 2.

Without elaborating this point, it should be sufficient to note in this connection the remarks of the Bengal Famine Inquiry Commission which pointed out that "the greater part of the population of Bengal did not suffer from lack of food in 1943."[1] The Commission further stated that after giving due weight to the greater difficulties faced by the Bengal Government, "we cannot avoid the conclusion that it lay in the power of the Government of Bengal by bold, resolute and well considered measures at the right time to have largely prevented the tragedy of the famine as it actually took place."[2]

6. TWO IMPORTANT MEASURES

Having satisfied ourselves about the interrelation of the Bengal Famine with the food resources of the country, it becomes necessary to examine the two important steps—the Grow More Food Campaign and Rationing—taken by the Government at that time. The Grow More Food Campaign was launched to save the cultivators of cash crops from ruin as well as to increase the production of food. The object of rationing on the other hand was to watch the interests of non-producers, particularly those who were employed in the essential services. As we would see many of the complications in the food economy of the country including increasing wheat imports in the decade that followed were directly connected with the latter step—rationing.

Grow more food campaign. The idea to launch the campaign was to increase production both by extensive and by intensive cultivation and the scheme was split up into various heads like reclamation of lands, diverting areas from cash to food crops, major and minor irrigation projects, land improvement schemes, manures and fertilizers, seed multiplication and distribution schemes, and provision of agricultural machinery and implements. The main success could, however, be achieved only in the field of diversion of land from short-staple cotton to foodgrains.

The campaign was launched early in 1942. As a result there was immediately a fall in the area under cotton. Nonetheless the area under rice and wheat did not show any appreciable increase. This was perhaps due to the cultivator's fear that any increase in the production of food will glut the market. This view is confirmed by the assurance given by the Government of India on May 20, 1942 that "should any development take place which will affect the saleability of the food crops, they will buy such quantities of food stuffs in the open market, whether in British India or the Indian States, as are calculated to prevent a serious fall in the prices."[3] When even this failed to instil confidence into the cultivators, the assurance was repeated by the Food Department on April 3, 1943 in a press note. Such an attitude on the part of the cultivators also proves incidentally the point that there was enough food in respect of demand in the country, otherwise they would not have missed the opportunity of producing more food on land than lying fallow. In spite of all this, the area under rice in India had been steadily increasing. It was this increase in the area under foodgrains that helped India to tide over the situation brought about by war.

---

[1] Report : *Op. Cit.* p. 68.

[2] Report : *Op. Cit.* p. 105.

[3] Quoted by Sir Henry Knight : *Food Administration*, 1954, p. 125.

Rationing. If India was not deficit in wheat, what made the imports of this commodity necessary ? An answer to this will perhaps be found in the scheme of food rationing which was introduced from early 1943.

Unprecedented transport bottlenecks which hampered the free flow of commodities from surplus to the deficit areas[1] and other wartime conditions helped to intensify local shortages of foodgrains.

Rationing was, therefore, introduced firstly to meet the demands of essential services and the non-producers and secondly to check inflationary prices. But the scheme involved certain commitments on the part of the Government for the fulfilment of which it was necessary to secure some stocks of foodgrains. It was this need which forced the Government to tap external sources when they failed to procure the necessary stock from the home market. Such steps which had already been taken by European countries like Germany and England[2] did not in any way indicate the existence of any shortages at home. Mr. W. H. Kirby, the British expert on rationing who had been invited by the then Government of India for advice had declared that "when supplies are ample, rationing is necessary for effective mobilization of resources against emergency.[3]

Procurement. With the introduction of food rationing, Government needed stocks to meet its commitments. Purchases from the open market were not possible under the conditions then prevailing in the country. A system of procuring foodgrains was accordingly considered necessary for the purpose. But even this scheme did not prove very successful owing to the non-cooperation of the various States—both surplus as well as deficit—with the Centre. Surplus provinces like the Punjab had a strong feeling that the other parts of India, specially the industrial areas, had unjustly been making a profit during the depression by living on cheap Punjab grains. They insisted on making good that loss and claimed special treatment in the matter of rationing.[4]

This was only one reason militating against the smooth working of the policy of procurement. The other as stated by Mr. Masani was that of defective policy of fixing price. "Procurement" in its whole history, according to him, "never reached the desired level."[5]

---

[1] S. G. Beri : *Basic Principles of Minimum Living Wage—A Review of Price Control in India*, 1943, p. 12.

[2] Germany did it in 1937. (Rationing and Control of Food supplies, J. M. Tinley, 1942, *Quarterly Journal of Economics*, May, 1941). Rationing in England was introduced on January 8, 1940. (*Food Rationing and Supplies*, 1944, League of Nations).

[3] Quoted by Rabindra Nath Chatterjee: *Price Control and Rationing in India*, March, 1943, p. 98.

[4] Sir Henry Knight : *Food Administration*, *Op. Cit.* p. 154.

[5] Legislative Assembly Debates, Vol. VII, 1946, p. 483.

7. THE BASIC PLAN

The Food Grains Policy Committee, 1943, recommended the formulation of a Basic Plan to assess the food requirements of the provinces. After the provinces declared whether they were surplus or deficit, the whole thing was thrashed out in discussions where efforts were made by the Central representative to secure as much as possible from the surplus states and supply as little as possible to the deficit ones. There was actually a sort of a tug of war between the various representatives each trying to secure the maximum for themselves. Table 1 gives the working of the Plan in so far as rice, wheat, millets and maize are concerned.[1]

TABLE 1. WORKING OF THE BASIC PLAN

year (1)	deficit declared by States (2)	surplus so declared (3)	gaps between 2 & 3 (4)	allotments made from the surpluses and imports (5)	quantities actually moved (6)
		all figures are in tons			
*1942–43	1,946	872	1,074	1,117	1,106
1943–44	4,995	2,236	2,759	2,757	2,584
1944–45	5,867	1,993	3,874	3,284	3,055
1945–46	5,081	1,784	3,297	3,221	2,972
1946–47	4,044	721	3,323	3,363	3,248

*Rice and millets only for 1942-43.

The gaps between deficit and surplus declared by the state under column 4 of Table 1 were the constant worry of the authorities and they were met through imports. As deficits by the provinces were normally inflated and surpluses deflated, these gaps continued to widen.

Dr. Rajendra Prasad,[2] the then Food Minister, while discussing the food situation in the country in the Legislative Assembly in 1946 made this point clear when he pointed out that the 'marketable surplus' in the country being 40 percent, procurements in the country (3.8 million tons) were only 16 percent of this surplus.[3]

Besides other war time commitments, the Government had to supply food to about 38 percent of the population. If all these commitments were to be honoured, it was necessary to procure the whole of the 'marketable' surplus. But the machinery for the procurement of foodgrains was in the hands of the provinces and full cooperation was not forthcoming from them.

Sir Henry Knight while describing the working of Basic Plan aptly remarked[4] that the Plan, owing to the selfishness of human nature and the constitutional handicaps of the

---

[1] Adapted from the *Food Administration* : *Op. Cit.* p. 183.

[2] Proceedings of the Legislative Assembly, November 4, 1946, Vol. VII, 1946, pp. 465-66 and 470.

[3] Sir Henry Knight : *Op. Cit.* p. 156, On difficulties in fixing targets.

[4] *Ibid*, p. 187.

Government of India in dealing with self-governing Provinces and States, met with many difficulties. The vicious circle started by rationing had permeated our body politic to such an extent that it became evident that imports of foodgrains were necessary to maintain the distribution system from toppling down.[1] The position became still worse after the war because of the inflationary movement which was in evidence from the second half of 1946. With the expectation of a sellers' market thus created, both the producer and the trader turned to hoarding in the hope of making profits.[2] The whole situation was further vitiated by black market prices. Such were the conditions which hampered a successful policy of procurement, thus necessitating increased imports.

Having examined the rationing and procurement systems as well as the working of the Basic Plan, we are now in a position to find an answer to the paradoxical situation of India's emergence as a net importer of wheat in increasing quantities. Out of the total foodgrains procured wheat procurements were the least and constituted hardly 10 to 15 percent; while rice procurement invariably represented more than 70 percent.[3] The only course left to meet wheat commitments under the circumstances was by way of imports.

As for the general position, the cultivator felt insecure and developed an apathy towards the city population in the words of Mr. Krishan Chandra Sharma. The conflict according to him was between the man who produces and the man who enjoys.[4]

8. THE PERIOD AFTER THE PARTITION OF INDIA

The year 1947 witnessed a new epoch in our food history. As a result of partition, the economy of the country, which was a homogenous unit, was disrupted. India got 562.6 million acres or nearly 83.6 percent of the total area leaving only 16.4 percent for Pakistan. At the time of partition, the total population of India and Pakistan was estimated at 417 million. After partition it was 337 million for India or nearly 80.2 percent of the total. The area between the two countries was thus divided practically in the same ratio as were their respective populations. But for a correct appraisal of the agricultural position of the country, it is necessary to consider also the factors like the irrigated area, the area actually under cultivation and such lands which can be brought under cultivation.

Net area sown and area under food. Out of a total net sown area of 286.6 million acres, the share of Pakistan was hardly 15.6 percent which may be increased by another 2 or 3 percent to account for the Pakistan States. The total was accordingly not more than 18 percent and even in the matter of total area under food crops, Pakistan did not get more than 19 percent. The position of Pakistan would thus seem to be in no way better than that of India in so far as area statistics are concerned.

Irrigated areas. Pakistan's immediate gain, however, was that she acquired 21.4 percent of the double cropped area and 29 percent of the irrigated area. This is the only factor which is said to be responsible for dislocating the food economy of India. But potential resources of India are more than sufficient to make up for this loss when it is realised that our irrigated area may touch the 100 million mark against the then 47.4 million acres when all our projects are completed.

---

[1] *Cf.* Also Dr. Anwar Iqbal Qureshi : *Op. Cit.* p. 6.

[2] *Cf.* Sir Henry Knight : *Op. Cit.* p. 267.

[3] *Indian Food Statistics*, 1949, pp. 32-35.

[4] Legislative Assembly Debates, Vol. VII, 1946, p. 486.

Culturable areas. The area of cultivable land in Pakistan is not much. Though we cannot lay our hands on any reliable data in this regard, available facts and figures are sufficient to prove our contention that the possibilities of extensive cultivation there are very meagre, while India has huge potentialities in this regard.[1]

Food production. An analysis of the total production of cereals will present practically the same picture. The Advisory Board of the Indian (then Imperial) Council of Agriculture Research (ICAR) estimated the average production of cereals in undivided India at 60 million tons. The share of the two dominions according to the Council, was in the ratio of 80 and 20[2] on the basis of average production figures. This ratio again tallies with that of the division of population and area between the two countries.

More than 90 percent of the population of Pakistan being rural, the consumption of cereals is also appreciably higher there. This is confirmed if we take the net availability of these commodities[3] as equal to the actual consumption levels in the areas concerned.

The study of agricultural economy in India before and after partition would be incomplete without a proper appreciation of the position in the princely States. At the time of partition, two powerful forces were at play, pulling in opposite directions. One was the disruptive force of communal exclusiveness which led to the secession from India of certain territories and their constitution into an independent State. The other was the revitalizing and uniting force of enlightened mutual interests in a free and resurgent India which swept away the barriers that separated the Indian States and Provinces.[4]

India was, as we have seen, a loser in the matter of irrigated and double cropped areas in so far as the Provinces were concerned. We have now to see how far this loss was made good by the integration of "Princely India." The position with regard to area and population of States before and after partition is given in Table 2.

TABLE 2. INDIA, PAKISTAN AND PRINCELY STATES

	United India	Indian Union	Pakistan	Total States	Indian States	Pakistan States
area (million acres)	1012	781	231	458	376	82
population 1941 (in millions)	389.0	318.9	70.1	93.2	89.0	4.2

Before partition, the total area of States was 45 percent of undivided India with only 24 percent population. A major portion of these States came over to India so that after partition, they constituted as much as 48 percent of the area of the Indian Union while their share of the population was only 28 percent. The contribution of these States to the net area sown was, however, only 25 percent of the total (60.8 million acres in a total of 241.9 million).

---

[1] P. C. Bansil : Extensive Cultivation, Rural India, September, 1956.

[2] Quoted by S. Thirumalai : *Post War Agricultural Problems and Policies in India*, 1954 p. 55.

[3] The Food Statistics : *Op. Cit.*, pp. 118 and 123.

[4] White Paper on States, 1950, p. 5.

It is true that much of the area in the States was underdeveloped, but it is difficult to believe that the percentage of uncultivated area was so high. A study of the reporting and non-reporting areas in the country, would also justify these doubts.

Out of a total of 781 million acres, returns from not more than 50 percent (377 million acres) were based on complete enumeration. The remaining 50 percent had as much as 224 million acres or 28 percent of the Union and 60 percent of the States areas as non-reporting. This in other words means that land utilisation returns were available from only 40 percent of the area of Indian States. Unclassified area in the Indian Union in 1949-50 was estimated at 196.2 million acres. Out of this 182.5 million acres (92.7 percent) was in the States.[1]

As against this, the only two States of any consequence which went over to Pakistan were Bahawalpur and Khairpur, a major portion of which was reporting. In the light of these facts, the view that the integration of States added to the responsibility of the Indian Union in the matter of food supply sounds utterly erroneous. More such discrepancies are now coming to light with the availability of improved data and the elimination of factors which were responsible for under estimation. Rajasthan, for example, has shown an increase of 15.3 million acres in the net area sown from 1948-49 to 1950-51. There was again a gap of the order 63 million acres in the area according to the Surveyor General and the Village papers. This gap in the year 1950-51 was reduced to a mere 7 million acres. Area under food alone increased by 7.2 million acres in 1953-54 over that of 1949-50. The Planning Commission in their Progress Report have accepted that this is mainly due to improved statistical coverage since 1949-50.[2] This is only one instance. The same is true in respect of other States which came over to India.

This was not all; even the yield per acre was underestimated to a great extent in these areas. Table 3 gives separately the estimated average yields per acre of all cereals (reporting areas only) in the Provinces and States separately from the year 1936-37 to 1947-48.[3]

TABLE 3. AVERAGE YIELDS OF FOODGRAINS IN PROVINCES AND STATES

period	provinces		states	
	yield in lb. per acre	percentage decrease	yield in lb. per acre	percentage decrease
1936–38	651	—	552	—
1938–42	639	2	515	7
1942–45	640	2	472	14
1945–48	594	9	327	41

As the areas in the States were practically interwoven with those of the Provinces, it seems rather incredulous that the yield per acre fell by 41 percent in the States as against 9 percent in the Provinces during the course of a decade. There is nothing to show that the soil in the States is much inferior as compared to that of the Provinces. But even if this hypothesis is accepted, this alone cannot explain such a heavy fall in the average yields in the States.

---

[1] Thirumalai : *Op. Cit.* p. 77,

[2] Five Year Plan Progress Report for 1953-54, p. 57.

[3] Adapted from Estimates of Area and Yield, *Op. Cit.* and Agricultural Situations in India for 1948–49.

Underestimation of the area sown in the States and the yields therefrom thus gave a wrong notion of the food position in the country. Otherwise it does not stand to reason that 48 percent of the land area could not feed even 28 percent of the population.[1]

This analysis helps us to understand two important points. Firstly, in the absence of reliable data from the State areas, it is not correct to assume that India suffered on the food front as a result of partition. Secondly, it would seem quite fallacious to think that the States were a liability on India in the matter of food.

Disturbances and food. We can thus conclude that as a result of partition of India there was no worsening of the food situation in the country in relation to its population. But communal disturbances, resulting in the movement of as many as about 12 million people on both sides of the border, disrupted the economy of the country as a whole.

There was a dislocation in occupations. Many of the agriculturist refugees from Lyallpur and Montgomery Districts of the Punjab, who had well irrigated lands in West Pakistan, felt disgruntled when suitable lands could not be made available to them immediately. There was, therefore, a tendency on the part of many to settle in urban areas. Most of the land, previously cultivated by Muslims, thus remained uncultivated.

"There has been a noticeable increase in the area of fallow land since the pre-war period, due perhaps in some measure, firstly, to restrictions on the cultivation of commercial crops and secondly to the dislocation in normal production caused by large scale movement of refugees after partition, unsettled conditions of tenure following the several reforms, which are in different stages of implementation, shortage of credit and supplies and adverse seasonal conditions, etc."[2] Table 4 gives the area under all cereals for three years during 1946-47 to 1948-49 and under wheat as well as other kharif cereals separately.

TABLE 4. AREA UNDER CEREALS 1946–1949

year	all cereals	wheat	kharif cereals*
	(million acres)		
1946–47	187.4	23.0	30.4
1947–48	169.0	19.4	13.9
1948–49	167.7	20.4	11.6

* Small millets only.

The area under cereals fell from 187.4 million acres in 1946-47 to 169.0 million acres in 1947-48. While the total area under cereals fell by 18.4 million acres in the post-partition year small millets alone suffered a reduction of nearly 16.5 million acres. The fact that the wheat acreage for 1948-49 had not fully recovered shows that the conditions had not become normal throughout the year 1948. The food shortage thus felt was purely temporary.

---

[1] *Cf.* Foodgrains Investigation Committee, 1950.

[2] *Agricultural Legislation in India*, Vol. III; *Agricultural Production and Development*, Directorate of Economics and Statistics, Ministry of Food & Agriculture, p. vi.

Increased rationing commitments. The situation became complicated in so far as Government commitments increased tremendously in 1947. Population under rationing swelled up from 1.6 million in 1943 to 148.7 million in March 1947. After Independence, population under statutory rationing went up by about 1.3 million, though there was an overall decrease in the rationed population because of a fall in the non-statutory area. But the Government commitments were enhanced by an increase in the area under statutory rationing, as an assured supply of foodgrains was required to be issued to every ration card holder in such areas.

The situation could only be tackled by the Government either by stepping up the procurement of food or by importing it. World food situation in 1947 was not very bright.[1] Balance of payment position was also adverse. There was an acute shortage of dollars, with the result that food imports fell to 2.6 million tons in 1947-48 from 3.1 million tons in 1946-47.

The procurement position was not better either. Figures for the three years 1946 to 1948 were as follows.

year	procurement (in thousand tons)
1946	3,091
1947	4,079
1948	2,650

With a fall in procurements as well as imports, and an increase in its commitments, the Government was in a very tight corner. Closing stock for the year 1947 was 645 thousand tons which were the least held in the entire period of rationing.

Decontrol. The Government faced a serious situation, popular opinion in the country was hostile to the imposition and continuance of controls. The campaign against them had, in fact, been launched much earlier by Mahatma Gandhi.[2]

It was decided on December 10, 1947, to embark on a policy of partial decontrol as recommended by the Foodgrains Policy Committee, 1948 which was set up for the purpose on September 27, 1947. The result was that the population under rationing came down from 147.9 million just before the introduction of decontrol policy in December 1947 to as low a figure as 71.5 million by the end of September 1948.

An interesting point to note in this connection is that these controls were done away with in a defeatist attitude.[3] When a country is under a controlled economy for a number of years, the success of such an experiment depends upon the ability of the Government to make available sufficient supplies to the non-producers. The stocks with the Government at that time were the least in the rationing history of the country. The necessary concomitants were thus conspicuous by their absence. Controls, bad as they were, could not be relaxed in such an unceremonial fashion.

---

[1] *Food Situation in India 1939-53* : Directorate of Economics and Statistics, 1954, pp. 168-69.

[2] *AICC Economic Review*, Vol. III, No. 1.

[3] There are opinions like that of N. V. Sovani (*Post-War Inflation in India—A Survey* 1948, p. 42) according to whom decontrol was launched in most favourable circumstances.

The country was suffering from a psychology of shortage and controls had become a vital factor in the economy. Political and communal disturbances, administrative chaos and allied factors added fuel to the fire. As a result of an over-burdened and ill-adjusted transport system[1] hoarding and black marketing, the prices rose sharply. These disruptive forces gained strength, because of the inability of the Government to meet the requirements of even those areas which normally depended upon the "marketable surplus." The vicious circle thus started gathered momentum and the Government had soon to retrace its steps.

Recontrol. The immediate result of decontrol was increased prices. The wholesale price index in the case of food articles shot up from 296.2 in September 1947 to 396.6 in September 1948. This rise of 34 percent in the case of food prices threw the Government in panic and the situation demanded immediate remedy. Controls were reimposed in September 1948.

Intensification of controls. The Government remained wedded to the policy of controls though the food situation in the country had become practically normal[2] in 1949-50.[3] Instead of bringing about an orderly reduction in their rationing commitments, the Central Government assumed supreme responsibility of feeding the deficit areas as well as a majority of non-producers. Population under rationing went up from 80.4 million at the end of 1948 to 119.6 million in 1949 and 124.5 million in 1950.

Procurement unrelated to production. The problem facing the Government under these circumstances was to establish a relation between the quantity of foodgrains procured and their production at home. In fact any increase in the production as a result of the 'Grow More Food Campaign' or the favourable weather was rarely reflected in increased procurements.

The production of foodgrains in 1950, for instance, touched the peak figure of 46.01 million tons as against 43.3 million tons during the previous year. The procurement of foodgrains, on the other hand, increased only by 2 lakh tons. Imports thus remained independent of increase or decrease in the production of foodgrains at home. This is explicit from a study of the Table[3] as well as the 1949 Policy Statement of the Government given below.[4]

"These latter commitments (refers to rationing) are met to a very large extent from imported grain. The quantity of imports is, therefore, not related to the total production of the country or any all-India production and consumption figures but is based on the specific demands of deficit Provinces and States for meeting their rationing commitments for a specific number of persons."

---

[1] *Cf.* Sovani, *ibid.*, p. 54.

[2] The Ministry of Food & Agriculture (Food Report for the year 1950-51 p.1).

[3] Govt. of India's policy statement dated July 1, 1949—The Times of India—July 19, 1949

[4] Policy Statement : *Op. Cit.*

TABLE 5. INTER-RELATION OF PRODUCTION, RATIONING (MILLION TONS) COMMITMENTS, PROCUREMENT AND IMPORTS

year	*population under rationing (millions)	production	procurement	percentage of 4 to 3	imports	total of procurements and imports	offtake
1	2	3	4	5	6	7	8
			(million tons)				
1947	144.7	42.544	4.08	9.6	2.33	6.41	7.067
1948	80.4	43.741	2.65	6.1	2.84	5.49	5.130
1949	119.6	43.314	4.38	10.1	3.71	8.09	7.692
1950	124.5	46.018	4.62	10.0	2.13	6.75	7.554
1951	122.0	41.744	3.77	9.0	4.73	8.50	7.867

*Figures relate to the population under rationing on December 31 of each year, except for 1947, in which case the date is December 1, 1947.

Sources : Bulletin on Food Statistics, January 1955.

Mr. K. M. Munshi, the then Food Minister, was of opinion that even self-sufficiency did not mean cessation of imports.[1] This shows that the production of the country was not allowed to play any part in the controlled economy of the country.[2]

Marketable surplus. That procurement was not keeping pace with production can also be proved from another angle. We have already referred to the remarks made by Dr. Rajendra Prasad about 'marketable surplus.' This in the normal pre-war days used to be of the order of 40 percent in the case of rice and 55 percent in the case of wheat.[3] It would on the other hand be seen from Table 5 (column 5) that procurement never exceeded the minimum of even 10 percent while the population under rationing remained in the neighbourhood of 30 percent.[4] The results of Diet Surveys conducted by the Indian Council of Medical Research for the period 1935-48, however, confirm that the consumption of cereals in the rural areas remained practically the same before and after 1944.

Again, there is possibility that the rural population, owing to increased prosperity, might have switched over to the consumption of fine grains. But factors like urbanisation and the recruitment of villagers in the army also released some of the foodgrains from the villages. Taking all these factors into account, it is reasonable to suppose that something like 40 percent of the total production of rice and wheat should have come up for the consumption of non-producers. Assuming that 10 percent out of this was retained for the rural non-producers, the balance could have been sufficient to meet the rationing commitments.

---

[1] K. M. Munshi : 'Our Food Objectives', *Illustrated Weekly*, August 12, 51.

[2] Mr. Munshi had himself said a year earlier (*Freedom from Foreign Bread*—p. 6) that "We shall make India self-sufficient by the end of 1951 so that we may be independent of foreign imports."

[3] Reports on the Marketing of Rice (p.7) and Wheat (p. 16)

[4] Dr. V. K. R. V. Rao also explained the position on the same lines at the 32nd Session of the Indian Economic Conference, 1950, *Indian Journal of Economics*, 1949-50, p. 233.

Procurement of foodgrains could not, however, reach the required levels. Firstly, because the pricing policy followed by the Government was faulty. While agricultural prices and subsidy schemes in other countries were so arranged as to pay higher prices on local production,[1] the position in India was just the reverse. Higher prices were being paid for the imported food and the home producer was asked to part with his produce at a much lower price. The Foodgrains Investigation Committee[2] also refers to the widespread dissatisfaction and discontentment among the producers of these rural areas which were under statutory rationing.

Secondly, the cultivator was subjected to many other unnecessary embarrassments. The Allahabad High Court held that the U.P. Foodgrains Procurement Order, 1948 went beyond the powers conferred by the Central Government.[3] The result of all this was that the position deteriorated further and imports instead of showing a fall went up. The solution thus lay either in stepping up procurement or in reducing commitments. The Government failed to pursue either course. They were unable to procure more food and even to abolish rationing. Such an action required boldness and imagination which the Government unforunately lacked at the time.

9. The planning commission and food policy

The Planning Commission also could not find out the real cause of increasing food imports before and during the war and post-war period. It recommended that food controls should be continued for the duration of the plan period and that imports should continue even if the target of 7.6 million tons of additional production was achieved. The Commission could never think in terms of self-sufficiency for India. Although it attached importance to the assessment of trends[4] in production and consumption, when laying out its final policy these two factors were totally ignored.[5]

Production depressed. Besides the inadequacy of procurements to meet the internal demand and a consequent increase in imports, controlled economy also had the effect of depressing home production in two ways.

In the first place, procurement prices were not fixed on the basis of any scientific principle.[6] They did not bear any relation to the cost of production nor had they any parity with the prices of industrial goods and raw materials like cloth and iron, which the cultivator had to purchase for his own requirements. Agriculture was never a paying industry,[7] but the cultivator was now made to feel that even when he had the chance of getting some remunerative prices, he was being deprived of his legitimate dues. All this had a depressing effect on the production of foodgrains. The cultivator was inclined to grow only such crops, as fetched him better prices.

---

[1] *Cf.* S. Kesva Iyengar : *Rural Economic Enquiries in the Hyderabad State Planning Commission*, 1949-50, p. 315.

[2] Report, p. 7.

[3] The judgement delivered by Aggarwal and Chaturvedi : *The Hindustan Times*, February 15, 1953.

[4] *The First Five Year Plan*, pp. 174-195.

[5] *The First Five Year Plan*, *Op. Cit.*, p. 175

[6] Unpublished M.A. thesis, P. C. Bansil : Feeding the Nation, pp. 63-68, for further reference.

[7] Dr. Baljit Singh : *Whither Agriculture in India*, pp. 194-196.

Secondly, owing to high prices of foodgrains in the foreign markets, the Government was obliged to subsidise imported grains. Large sums of money were spent to provide cheap food in the deficit areas. This created the impression among the farmers of "surplus" areas that they were being ignored and neglected. Their own surplus produce was being procured at cheaper rates. High price was being paid for foreign imports and a sizeable sum of money was being spent in the deficit areas on food subsidies. This created a feeling that one section of the population was being fed at the cost of the other. Cultivators were, therefore discouraged and they lost whatever little incentive they had to produce.

The food problem was in this stalemate in the closing days of 1951 and early in 1952. The self-sufficiency pledge had not been fulfilled. The stress on food imports seemed to be more pronounced. The import target originally fixed for the year 1952 was in the neighbourhood of 7 million tons. It was subsequently reduced to about 5 million tons. Even then this figure was higher than the previous year's record imports. No solution seemed to be in sight. This was, perhaps, the worst phase in the food history of India. The climax was reached and a new chapter in the food debacle was in the offing.

Change in food policy. The year 1952 opened with a very gloomy picture on the food front. The withdrawal of food subsidies in March, 1952, aggravated the tense atmosphere in the cities. The country was considered far from self-sufficient in food and there was no question of reducing food imports.

In May of the same year the late Rafi Ahmed Kidwai took charge of the Food Ministry. He undertook a hurried survey of the affected areas and made a categorical statement that there was no shortage of food in the country. According to him if at all there was any problem, it was one of purchasing power. Once this was realised, the next step —relaxations of controls—was but a simple process.

Madras experiment. The Planning Commission did not view with favour the idea of relaxing controls. They thought it might interfere in the planning which to a large extent was based on deficit financing. But R. A. Kidwai succeeded in finding an ally in C. Rajagopalachari, who introduced the policy of decontrol in Madras—a highly deficit State—in June, 1952. The whole of the State was divided into six zones allowing movement of rice within the zone. This sets in motion the strange phenomenon of rising prices in the surplus districts and of falling prices in the deficit ones. The critics raised a hue and cry and quoted the example of the 1947 experiment. The trend in the deficit districts was, however, reversed in the beginning of August and with the coming in of new crops in September, prices began to fall throughout the province.

This strengthened the hands of Kidwai and put him in a position to meet any challenge. Gradually all the States were persuaded to follow the policy of partial decontrol, with the result that within about 2 years, practically all the irksome controls on food were removed and cries of surplus food were heard from almost all parts of the country.

Working of decontrol. For a correct assessment of the country's food position, it is essential that we study thoroughly the decontrol experiment. Was a reduction in imports due to the return of normal conditions—stocks in the surplus areas supplying the needs of the deficit areas—or were the non-producers placed at the mercy of the producers?[1] A reply to this question would provide the real answer to the problem at hand.

The immediate step taken in derationed areas was that ration shops were converted into fair price shops. The card holders in the original statutory rationed areas were allowed to draw their ration as usual from these fair price shops, if they so liked. The only change was that the free movement of foodgrains within the areas was permitted so that the forces of demand and supply could control the market. The number of such fair price shops was determined by the needs of a particular area. Data regarding changes in the number of fair price shops and their monthly off-take of foodgrains are not available for all the States. Table 6, however, gives such figures in respect of Madras, Bihar and Uttar Pradesh.[2]

TABLE 6. FAIR PRICE SHOPS AND THEIR MONTHLY OFF-TAKE

year/month	Madras		Bihar		Uttar Pradesh	
	fair price shops (in thousand)	off-take (in thousand mds)	fair price shops (in thousand)	off-take (in thousand mds)	fair price shops (in thousand)	off-take (in thousand mds)
1952						
June	—	74	10.9	15	—	67
September	12.2	147	—	24	2.3	76
December	—	22	—	10	—	41
1953	—					
March	—	34	—	8	—	43
June	9.9	58	3.9	9	1.1	4

It would be seen that the number of fair price shops was governed by the rationed off-take. It was increased or decreased corresponding to an increase or decrease in off-take of foodgrains so that the interests of the consumer were not sacrificed at any time. In Bihar, for example, when the rationed off-take fell from 15 thousand tons in June, 1952 to 9 thousand tons in June 1953, the number of fair price shops was reduced from 10,953 at the end of June, 1952 to 3,884 at the end of June, 1953. Same was the case in other States.

In West Bengal where controls were relaxed in January 1953, the number of fair price shops was reduced from 3,218 at the end of December, 1952 to 225 at the end of March, 1953. When, however, market prices showed a tendency to rise in May, 1953, the number of fair price shops was increased to 2,030 at the end of June, 1953 so that by December, 1953 prices of rice came down to those of the corresponding period in 1952 as would be clear from Table 7 below.

---

[1] *Cf.* The controversy over 'Why Decontrol' ? by P. C. Bansil and Prof. Hartirath Singh in the Hindustan Standard, Economic Supplement, dated Sep. 28, 1953 and Nov. 23, 1953. And an article On Our Food Policy, P. C. Bansil : *Economic Weekly*, Bombay, October, 24, 1953.

[2] Adapted from Agricultural Situation, August 1953, pp. 338 & 339.

TABLE 7. PRICE OF RICE AT SELECTED CENTRES (WEST BENGAL)

centre	1952	1953		
	December	March	June	December
		(in rupees per maund)		
Contai	24-	16-	21-	12-14-
Bolpur	14-	16-	22-8-	14- 4-

Every possible effort was thus made to see that there was no rise in the price of foodgrains after the relaxation of controls.[1] In spite of that there was a general rise in prices immediately after decontrol so that Economic Adviser's Index for cereals stood at 480.0 in 1953 as compared with 445.0 in June 1953. But even then the prices were lower than those ruling in the open market during the corresponding period of the previous year. Again, the percentage rise in the case of cereals was only 7.8 as against 16.41 for food articles. This is illustrated in Table 8.

TABLE 8. PRICE INDEX—WITH BASE, SEPTEMBER 1939 AS 100

	June 1952	August 1953	percentage rise
food articles	349.6	407	16.41
cereals	445.0	480	7.8
pulses	434.0	504	16.1
industrial raw materials	430.3	488.7	13.67
all commodities	374.8	410.4	9.3

In early 1954, food prices started falling below uneconomic levels, with the result that the Government had to announce the policy of price support, for cereals and floor prices were fixed. In the case of wheat, it was Rs. 10 per maund. Not only that, efforts were made to find export markets for rice, a commodity which had been imported for many years in the past.

The decontrol policy reached its climax on July 9, 1954 when all controls on rice were also removed. This in spite of a shortfall of the order of 3 million tons in the rice crop of 1954-55 when compared with the preceding year, led to further easing of the situation from the supply side. Prices continued to have practically the same trend with the result that the Government was faced with the problem of extending the price support policies to rice as well as other coarse grains like jowar, bajra and gram.

---

[1] For the various measures adopted by the Government refer to Agricultural Situation in India, August 1953, p. 344.

It may be argued that good agricultural seasons resulting in bumper crops for the consecutive years 1952-53 and 1953-54 accounted for this bright picture on the food front. There is no denying the fact that nature was exceptionally kind during these two years. Cereal production in the year 1953-54 was 57.9 million tons—nearly 12 million tons above that of 1949-50 and 9 million tons over that of 1952-53.

It does not, however, appeal to reason that just two or three good monsoons were able to wipe off the supposed deficit in foodgrains which is said to have existed over such a long time. What seems plausible is that a major portion of this so-called increase was merely statistical increase.[1]

One such case was that of Rajasthan. The position was not otherwise different in other Part 'B' States where there was hardly any machinery for the collection of agricultural data in the past. Out of the total increase of 12 million tons as much as about 5 million tons or 41 percent is accounted for by Part 'B' States. If the season was good for these areas, it was equally good for other parts of the country where increases in food production were not so marked. It is also to be noted that much effort has not been made to increase agricultural production by way of irrigation facilities, provision of better seeds etc. in these backward areas by the Planning Commission under the 'Grow More Food Campaign.' A sizeable portion of this increase was thus the one which had not originally been accounted for.

Nature was not so kind in the year 1954-55 when cereal production fell down by about 3 million tons to 55.3 million tons. The floods of 1955 were said to be the worst over the last century. The food position in the country, however, remained as bright as before. This being the case, we would be justified to conclude that the bright position on the food front during the period was not due merely to better season or even special efforts made by the Planning Commission.

The adverse season of 1954-55 when there were floods over an area of some 25,650 square miles affecting some 95 lakhs of people in Assam, West Bengal, Bihar and the United Provinces alone, also did not bring any change for the worse. The floods again devastated vast areas in 1955-56 in the affected territories. The Prime Minister while speaking about the United Provinces said that "no such floods have taken place in U.P. during the last 85 years." The loss suffered by the Punjab and PEPSU was the worst in their histories. The food position, however, remained quite easy.

After an easy position, for about 2 years, the last two years—1956 and 1957—have taken a turn for the worse, against an all-time record production of 28.1 and 9.1 million tons of rice and wheat respectively, for the year 1956-57. This brings to focus the point that the supply of foodgrains is not the same as production and demand is not the same thing as consumption. Even if prices are assumed to be determined by the interaction of supply and demand, in the case of foodgrains supply is invariably determined by the quantity released by the producers, technically known as the 'marketable surplus' and demand is not that of the producers, but only that of the non-producers.

---

[1] P. C. Bansil: Food Statistics, *Sankhyā*, **17**, 247-250, (1956).

Due to rapid industrialisation and many other factors, the number of non-producers is increasing rapidly. So much so that the increase in urban population is expected to grow at an annual rate of 4.8 percent during the second plan period. It is quite possible that with the increasing emphasis on cottage and small scale industries, the non-producers in the rural sector may also increase. The effective demand for foodgrains, in the restricted sense explained above, has thus increased appreciably. This would mean that unless effective supply, i.e., marketable surplus, shows a corresponding increase, there is bound to be a disequilibrium in the forces of effective demand and effective supply. If this disequilibrium is fed on certain external forces, militating against them, the normal forces of supply and demand are liable to be disturbed further, either bringing a swing to the downward or upward.

Immediately after decontrol in 1952-53 and 1953-54, a crash in the world commodity market, the dehoarding spree and exceptionally good harvest of 1953-54 at home, can easily be considered as the factors responsible for the continued fall in food prices right upto the early part of 1955. The turn came in the later part of 1955. The immediate responsibility for this may be attributed mainly to psychological factors like failure of crops in many parts of the country, decision to spend Rs. 7200 crores on the second five year plan, with a big unfilled gap for deficit financing and as much as Rs. 6200 crores representing investment, i.e., expenditure on the building up of productive investment. Once the market psychology is changed, the usual phenomena of unplanned movement of foodgrains over long distances in search of abnormal profits, transport bottlenecks, and hoarding at all the possible levels in the hope of getting more profits, could easily be witnessed in the foodgrains markets.

The immediate step necessary was the restrictions on the movement of foodgrains over long distances so as to relieve pressure on transport. But the zonal scheme, creating three wheat zones and one that for rice came as late as July, 1957. While the pressure on demand during all these days was increasing due to factors explained above, the marketable surplus was being adversely affected. In the earlier period, it was only the trader and the middle man who were responsible for speculative hoarding, but the imperceptible effect of recent changes in the land policy, increases in the income of the cultivator and his consequent ability to improve his retentive capacity, have brought on the scene another class which has aggravated what is known as the propensity to hoard.[1] Such small changes in the holdings of stocks, tend to modify significantly the marketable surplus, resulting in an aggravation of food prices.

The only remedy at the moment would thus seem to be the building up of market psychology by an enhancement of the supply. This is possible by the release of food stocks in response to genuine or even ingenuine demands. It can be done either by the purchase of foodgrains from the open market or imports from abroad. Since the propensity to stock at the moment is the maximum, it would perhaps not be advisable to resort to any large purchases from the home market. The present policy of importing foodgrains, for the building up of reserves would seem to be the most sound one, under the circumstances, the other suggestion of the Foodgrains Enquiry Committee regarding the establishment of regulated markets[1] for encouraging the flow of marketable surplus also deserves consideration.

---

[1] *Cf.* Foodgrains Enquiry Committee, pp. 44-46 for similar details.

Once market psychology is thus built up, the forces responsible for the propensity to stock are bound to get a setback and a tendency to dehoard is likely to come into motion. Other palliative measures like credit restrictions, powers to freeze stocks under the Essential Supplies Act, and Zonal restrictions, would all the same be necessary for so long as the country is not able to have an effective hold on natural hazards like floods and droughts by a systematic irrigation system.

This does not mean that everything is allright on the production side. The population of the country is increasing steadily at an annual rate of 4 to 5 millions and additional food will be required to feed the new mouths.

The production of food, as that of any other agricultural or industrial product, will have to conform to objectives of a planned economy and thus help in the realisation of our dream of a better and happier India.

Food potential of the country being rather high, there is no danger of a long period problem. The possibility of that of a short period one cannot, however, be ruled out. This is because

(a) Indian agriculture is still subject to hazards of nature, and every two years, in a cycle of five are normally bad years.

(b) An absolute increase in production has no significance in a developing economy, if the non-producers fail to get the required share of their demand through the normal functioning of the market. What is known as 'Scissors Crisis' in Russian history is well known.

As regards (a) the remedy lies in increasing production in order to have surplus food in good years so as to cover the deficit supply in years of bad harvests. This should not be difficult if the newly constructed irrigation works are fully utilised and deterioration of land from old ones is protected. The Deputy Chairman Planning Commission in his recent address to the Development Commissioners at Mussorie, has brought to light the sad plight of the States having utilized hardly 4 million acres out of the 6 million acres for which irrigation facilities were provided during the first five year plan period. Similarly an expert team on plan projects under the chairmanship of Shri Gadgil has revealed in its draft report that several million acres of the most fertile land in the Punjab have been ruined by water-logging and alkalinity largely as a result of seepage of water from canals and distributaries. If the total food potential of the country is to be exploited, all such matters will need due attention.

As for (b), the existing stocks will have to be properly husbanded. The marketable surplus can be reduced either by increasing consumption on the farm or by some impediments being placed in the way of normal trade channels; may be due to transport bottlenecks or hoarding at some of the intermediary stage. The first and foremost condition to overcome such a situation is the building of reserve stocks which must be unloaded at the appropriate time with a proper understanding of the market psychology. More so, when a system of controlled supply of foodgrains to the non-producer is not only difficult, but full of still greater dangers. Whether these food reserves are built by purchases from open market, imports or even conversion of the present system of land revenue in cash to that in kind as is the case in China, needs a thorough examination, which is beyond the scope of the present study.

---

[1] *Ibid.*, p. 119.

The Produce (Development and Warehousing) Corporation Bill, 1946, was timely from this point of view. Besides its functions of storage and warehousing, the Central National Co-operative Development and Warehousing Board is planning to formulate programmes for the production, processing, marketing, imports and exports of agricultural produce with the help of co-operative societies.

Credit control, restricted money supply and changes in the reserve ratio under section 42 of the Reserve Bank Act would also be necessary to curb speculative activities. It may become imperative at times to have control by fixing ceiling prices rather than by state trading. Such a measure on an all-India basis would be quite feasible. When the Centre has already acquired such powers by an amendment of the Constitution to have control over the Essential Supplies Act. In order to make this policy of price control a success, a scheme of fair price shops, particularly in the deficit pockets would be quite useful. They will help to build up market confidence and serve as an effective check on hoarding as well as spiralling prices.

Most of the transport bottlenecks will have to be removed by an improvement of the governmental transport means. But in the short period, the creation of food zones as constituted at present, would serve the purpose. Such a system would obviate the haphazard and unplanned movement of foodgrains over long distances which is primarily motivated by private profits rather than serving the national interests. A zonal scheme if judiciously formed and fully implemented, is sure to bring out hoarded stocks in due course; thus adding to the marketable surplus.

There is also the need to keep a watch on food subsidies. Food imports even today are being heavily subsidised. Such subsidies help to keep up consumption at artificial levels. They are a heavy drain on the exchequer. The whole policy deserves a careful re-examination. The earlier they are removed, the better it would be for the country.

Least, but certainly not the least, a well planned policy of a change in the existing food habits would not only alleviate the short period difficulties from time to time, but also help to improve the quality of food. A shift from rice to wheat, gram and millets, for instance, would be quite welcome to all students of human nutrition. Some of the subsidiary foods like potato can also play a major role to fill the gaps in periods of stress. If short period problems are adequately tackled in the manner described above, the long period, it can be emphasised, will take care of itself.

Paper received : June, 1958.

TEST SCORING BY THE 602A CALCULATING PUNCH

By AJIT HALDAR

Indian Statistical Institute, Calcutta

SUMMARY. A procedure for test scoring with the help of 602A Calculating Punch is described in this paper.

Leonard Staugas (1954) has given a method using the ESM, for the same purpose. Although the 602A Punch is a slower machine compared to the ESM, it is hoped that in installations where the ESM is not available, or under other conditions in which the 602A may be preferred, that the present method would be of value. Also, when the number of candidates and tests is small (e.g., when the total number of cards involved is less than 2000) the relative handling time becomes lower than in the ESM method.

Statement of the problem. Suppose for a particular test there are 50 questions. Each question has several alternative answers numbering 1 through 5 of which only one is correct. An examinee just ticks off the answer for each question. The answers (i.e., the numbers ticked off) are then transferred into punched cards according to a suitable design. The ultimate end is to find out the total number of right scores for each candidate.

Machine Operations. The detailed steps are stated below :

(1) *Card design.* Individual cards are punched for each candidate, according to the following card design :

CARD DESIGN

sl. no.	description		card column	remark
1.	candidate number		1 — 3	
2.	test number		4 — 5	
3.	question 1		6	
.	. . .		.	
.	. . .		.	
.	. . .		.	
52.	question 50		55	
53.	total of the scores for questions	1-10(=a)	56—57	to be punched by 602A
54.	,, ,, ,,	11-20(=b)	58—59	
55.	,, ,, ,,	21-30(=c)	60—61	
56.	,, ,, ,,	31-40(=d)	62—63	
57.	,, ,, ,,	41-50(=e)	64—65	
58.	grand total ($a+b+c+d+e$)		66—67	

For convenience of drawing the wiring diagram, the following arrangement of questions are taken.

Questions	Code for correct answer
1–10	1
11–20	2
21–30	3
31–40	4
41–50	5

(2) *Computation on* 602*A*. The method of wiring is as follows :

Read Cycle. The columns for which code for correct answer is '1' (Q1-10), are entered in storages, 1L, 1R, 2L, 2R, 3L taking only two columns in one storage. These storages are impulsed to "Read in".

Programmes 1 *to* 5. Through jack plugging storage, R.O. hubs are connected to common hubs of two digit selectors. Digit selector '1' hubs are connected to immediate pick up of pilot selectors 1 and 2. Then common hubs of corresponding pilot selectors are connected to digit emitter '1' and transfer to counters 1 and 2 respectively. At subsequent programmes, storages are read out but only one at a time and this is continued until all the storages are exhausted. Counters 1 and 2 are impulsed to add at each programme step.

Programmes 6 *and* 7. Cross-footing and punching of a are done in programmes 6-7. The wiring diagram for punching a, which will be of interest is given below :

The procedure is repeated for calculating b, c, d and e. Corresponding digit hubs from digit selectors are connected to P.U. hubs of pilot selectors. A final cross-footing gives the total score $(a+b+c+d+e)$.

Remark. A model case consisting of 50 questions and answer codes 1 through 5 has been dealt with here. The number of questions or codes may be different in different problems. It is easily seen how the plugging will then have to be altered accordingly.

Acknowledgement. My thanks are due to the technicians of the Psychometric Research and Service Unit of the Indian Statistical Institute for giving me the problem and also to my colleague Shri Hiren Das for valuable suggestions in preparing the paper.

Reference

Staugas, Leonard (1954) : A rapid method for scoring tests punched in IBM cards. *Educational and Psychological Measurement*, **14**, 1, pp. 101–105.

Paper received : April, 1958.

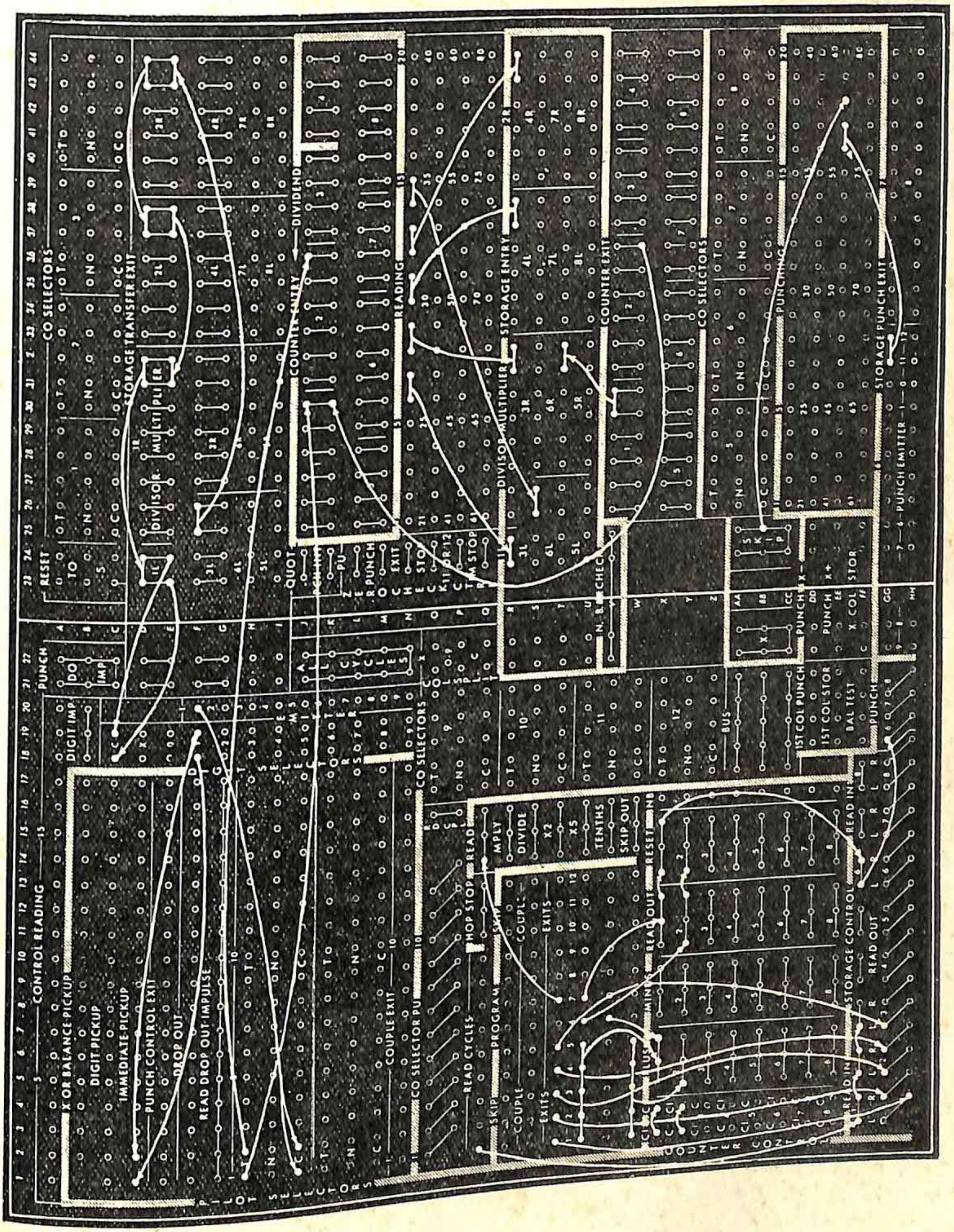


Figure 1. Wiring Diagram

CORRIGENDA

Unbiassed Minimum Variance Estimation in a Class of Discrete Distributions : By J. Roy and S. K. Mitra, *Sankhyā*, **18,** 376-378.

The following corrections are to be made in Table 2.

(t, n)	Incorrect	Correct
9,6	0.950	0.905
11,9	0.419	0.429
12,8	0.986	0.896
20,9	1.096	1.906
30,9	3.121	3.212
32,9	3.334	3.455
32,10	2.950	3.062
44,9	4.957	4.857
53,9	5.867	5.876
71,8	8.784	8.874
76,10	7.579	7.597

Almost Unbiased Ratio Estimates Based on Interpenetrating Sub-Sample Estimates : By M. N. Murthy and N. S. Nanjamma, *Sankhyā*, **21**, 381-392

Equation (4.2) : $n^{3/2}$, outside the curly bracket in the numerator instead of n.

Footnote line 4, p. 385 : $(v(\hat{x})+nx^2)^3$ instead of $(v(\hat{x})+nx^2)$ in the denominator of the second term.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Sri N. T. Mathew
Dr. A. Matthai
Dr. Sujit Kumar Mitra
Professor M. Mukherjee
Sri P. Pant
Dr. C. R. Rao
Dr. J. Roy

Editor,
Sankhyā: The Indian Journal of Statistics

SANKHYĀ

THE INDIAN JOURNAL OF STATISTICS

Edited by : P. C. MAHALANOBIS

VOL. 22, PARTS 3 & 4 JUNE 1960

THE ADDITION OF RANDOM VECTORS

By J. B. S. HALDANE

Indian Statistical Institute, Calcutta

SUMMARY. Expressions are obtained for the resultant of n vectors whose directions are completely random in s dimensions, but for each of which the distribution of the scalar square is given. This is most readily calculated from the vectorial cumulants, which are defined, and shown, besides their cumulative property, to measure departure from a gamma distribution.

INTRODUCTION

I consider the addition, or composition, of vectors whose direction is completely random in s dimensions. Randomness of direction is a very much simpler notion than randomness of a scalar number, as appears from the various formulations which Jeffreys (1933) has used to express his ignorance of the value of a scalar. By the statement that the direction of a displacement is completely random I mean that if a point is displaced through a distance x the probability that it will be displaced into any region of equal magnitude (length, area, volume, etc.) of the $(s-1)$ dimensional variety (circle, sphere, etc.) consisting of points at a distance x in Euclidean or other isotropic space from its original position, is equal. In particular the probability that a one dimensional vector will have a positive or negative value must be equal. When we have said that the direction is completely random we know nothing about the scalar magnitude. I shall suppose that the distribution of its square, that is to say of the squared magnitude of a displacement, or the scalar square of a vector, is specified by its cumulants κ_r. We might have used the moments μ_r of the square, or the even cumulants κ'_{2r} or moments μ'_{2r} of the displacement, but these lead to less compact expressions. I shall later introduce the vectorial cumulants θ_r which have several properties analogous to cumulants in one dimension. As however they are functions of the number s of dimensions it is best to show how they arise naturally in the course of the analysis.

It is worth while considering for a moment what happens if the directions of the vectors added are incompletely random. The simplest types of departure from complete randomness are of two kinds. The directions of all vectors may be unequally frequent in different directions, for example displacements approximately north or south may be commoner than those approximately east or west. Unless the probabilities of displacements in two opposite directions are equal there is no advantage in using scalar squares. But even if they are, it is often best to consider displacements in mutually orthogonal directions separately. Secondly the angle between two successively added vectors may have some other distribution than that appropriate to s dimensions. If so, addition is not in general associative, though we shall discover the conditions under which it is so. Here however there is no advantage in resolving the vectors orthogonally.

I have deliberately avoided such resolution because in the data which induced me (Haldane, 1960) to undertake this investigation, namely those of Alström (1958), there was a considerable positive correlation between the squares of displacements in orthogonal directions. The most general problem of the addition of random vectors would, I suppose, be something like this. We have added a series of vectors $[x_1, y_1, z_1 \ldots], [x_2, y_2, z_2 \ldots], \ldots [x_{n-1}, y_{n-1}, z_{n-1}, \ldots]$ in that order, x, y, z, etc. being resolutes in s mutually orthogonal directions. Given these vectors we can predict the frequency with which the resolutes $x_n, y_n, z_n, \ldots$ of the n-th vector lie between infinitesimal limits. We have to determine the distribution of the sum of these n vectors. If there are s dimensions then in the most general case the s scalars $x_n, y_n, z_n \ldots$ have a joint distribution whose parameters depend on the $s(n-1)$ components of the preceding vectors, and on an indefinite number of arbitrary constants. The problem is hopelessly complicated. In what follows I assume that successive vectors are independent as regards their magnitude, though the distribution of the angle between them, which plays a part comparable to correlation, depends on the number of dimensions.

THE ADDITION OF TWO RANDOM VECTORS

Let **X**, **Y**, be two random vectors in Euclidean space. Let the non-negative numbers x and y be their scalar squares $\mathbf{X} . \mathbf{X}$ and $\mathbf{Y} . \mathbf{Y}$, and let ϕ be the angle between their directions. As the resolutes of the vectors may be negative we can suppose $0 \leqslant \phi \leqslant \frac{1}{2}\pi$. If we assumed $0 \leqslant \phi \leqslant \pi$ we should have to add that the distribution function of $\phi - \frac{1}{2}\pi$ was even, otherwise we could not operate with squares. Then if **Z** is the sum (resultant) of **X** and **Y**, and $z = \mathbf{Z} . \mathbf{Z}$ is the scalar square of **Z**,

$$z = x + y \pm 2\sqrt{xy} \cos \phi. \qquad \ldots \quad (1)$$

Let κ_r, λ_r, ζ_r be the r-th cumulants of the distributions of x, y, and z. I assume that positive and negative values of the ambiguity are equiprobable, so that the expectations of odd powers of $\pm\sqrt{xy} \cos \phi$ are zero. Let

$$\mathscr{E}(\cos^{2r}\phi) = c_{2r}.$$

The distribution function of ϕ in s dimensions is

$$dF \propto \sin^{s-2}\phi \, d\phi \qquad (s > 1).$$

Thus

$$c_{2r} = \int_0^{\frac{\pi}{2}} \sin^{s-2}\phi \cos^{2r}\phi \, d\phi \div \int_0^{\frac{\pi}{2}} \sin^{s-2}\phi \, d\phi.$$

Hence $c_2 = \dfrac{1}{s}$, $c_4 = \dfrac{1.3}{s(s+2)}$, $c_6 = \dfrac{1.3.5}{s(s+2)(s+4)}$, $c_{2r} = \dfrac{1.3.5\cdots(2r-1)}{s(s+2)(s+4)\cdots(s+2r)}$. ... (2)

So

$$\mathcal{E}(z^r) = \mathcal{E}\left[(x+y)^r + 4\binom{r}{2} c_2 (x+y)^{r-2} xy + 16 \binom{r}{4} c_4 (x+y)^{r-4} (xy)^2 + \cdots \right]. \quad \ldots (3)$$

This is of course a terminating series with $\frac{1}{2}r+1$ terms if r is even, and $\frac{1}{2}(r+1)$ if r is odd. Before calculating the values of ζ_r it is instructive to consider the cases when $s = 1$ and $s = \infty$. The latter is very simple. As s tends to infinity almost all angles between **X** and **Y** are right angles, and their scalar squares become additive, so that $z = x+y$, and $\zeta_r = \kappa_r + \lambda_r$, the well-known additive property of cumulants. When $s = 1$, so that $\phi = 0$, the even cumulants of $x^{\frac{1}{2}}$ have this property. If κ'_{2r} be the $2r$-th cumulant of the distribution of $x^{\frac{1}{2}}$ we write $\kappa'_{2r} = \theta_r$, and find, inverting equations (3) of Haldane (1941), then :

$$\begin{aligned}
\theta_1 &= \kappa_1 \\
\theta_2 &= \kappa_2 - 2\kappa_1^2 \\
\theta_3 &= \kappa_3 - 12\kappa_1\kappa_2 + 16\kappa_1^3 \\
\theta_4 &= \kappa_4 - 24\kappa_1\kappa_3 - 32\kappa_2^2 + 272\kappa_1^2\kappa_1 - 272\kappa_1^4 \\
\theta_5 &= \kappa_5 - 40\kappa_1\kappa_4 - 200\kappa_2\kappa_3 + 880\kappa_1^2\kappa_3 + 2{,}400\kappa_1\kappa_2^2 - 9{,}920\kappa_1^3\kappa_2 + 7{,}936\kappa_1^5 \\
\theta_6 &= \kappa_6 - 60\kappa_1\kappa_5 - 480\kappa_2\kappa_4 - 452\kappa_3^2 + 4{,}200\kappa_1^2\kappa_4 + 22{,}368\kappa_1\kappa_2\kappa_3 + 15{,}360\kappa_2^3 - \\
&\quad -29{,}344\kappa_1^3\kappa_3 - 496{,}128\kappa_1^2\kappa_2^2 + 1{,}364{,}528\kappa_1^4\kappa_2 - 1{,}439{,}552\kappa_1^6. \qquad \ldots (4)
\end{aligned}$$

The values of κ_r in terms of θ_r, θ_{r-1}, etc, have already been given by Haldane (1941). The expressions (4) are of no value in this investigation, except as special cases of the expressions (9) derived later which hold in any number of dimensions, on which they furnish a useful check.

From (3) we can write down expressions for the expectations of powers of z, and hence for the cumulants of its distribution. If κ_r and λ_r are the r-th cumulants of the distributions of x and y, whilst ζ_r are those of z, we find :

$$\begin{aligned}
\zeta_1 &= \kappa_1 + \lambda_1 \\
\zeta_2 &= \kappa_2 + \lambda_2 + 4c_2\kappa_1\lambda_1 \\
\zeta_3 &= \kappa_3 + \lambda_3 + 12c_2(\kappa_1\lambda_2 + \lambda_1\kappa_2) \\
\zeta_4 &= \kappa_4 + \lambda_4 + 24c_2(\kappa_1\lambda_3 + \lambda_1\kappa_3 + 2\kappa_2\lambda_2) + 16c_4(\kappa_2 + \kappa_1^2)(\lambda_2 + \lambda_1^2) - 48c_2^2\kappa_1^2\lambda_1^2 \\
\zeta_5 &= \kappa_5 + \lambda_5 + 40c_2[\kappa_1\lambda_4 + \lambda_1\kappa_4 + 3(\kappa_2\lambda_3 + \lambda_2\kappa_3)] + 80c_4[\kappa_2\lambda_3 + \lambda_2\kappa_3 + \kappa_1^2\lambda_3 + \\
&\quad + \lambda_1^2\kappa_3 + 2(\kappa_1 + \lambda_1)\kappa_2\lambda_2 + 2\kappa_1\lambda_1(\kappa_1\lambda_2 + \lambda_1\kappa_2)] - 480c_2^2\kappa_1\lambda_1(\kappa_1\lambda_2 + \lambda_1\kappa_2) \ \ldots \qquad (5)
\end{aligned}$$

The expression for ζ_6 is considerably more complicated, containing terms in c_2^3, $c_2\,c_4$, and c_6. I do not think it is likely to be required in practice. Inserting the values of c_2 and c_4 given by (2) we have :

$$\begin{aligned}
\zeta_1 &= \kappa_1+\lambda_1 \\
\zeta_2 &= \kappa_2+\lambda_2+4s^{-1}\kappa_1\lambda_1 \\
\zeta_3 &= \kappa_3+\lambda_3+12s^{-1}(\kappa_1\lambda_2+\lambda_1\kappa_2) \\
\zeta_4 &= \kappa_4+\lambda_4+24s^{-1}(\kappa_1\lambda_3+\lambda_1\kappa_3)+48s^{-1}(s+2)^{-1}(s+3)\kappa_2\lambda_2+48s^{-1}(s+2)^{-1} \\
&\qquad (\kappa_1^2\lambda_2+\lambda_1^2\kappa_2)-96s^{-2}(s+2)^{-1}\kappa_1^2\lambda_1^2 \\
\zeta_5 &= \kappa_5+\lambda_5+40s^{-1}(\kappa_1\lambda_4+\lambda_1\kappa_4)+120s^{-1}(s+2)^{-1}(s+4)(\kappa_2\lambda_3+\lambda_2\kappa_3)+ \\
&\qquad +240s^{-1}(s+2)^{-1}(\kappa_1^2\lambda_3+\lambda_1^2\kappa_3)+480s^{-1}(s+2)^{-1}(\kappa_1+\lambda_1)\kappa_2\lambda_2-960s^{-2}(s+2)^{-1} \\
&\qquad \kappa_1\lambda_1(\kappa_1\lambda_2+\lambda_1\kappa_2). \quad \ldots \quad (6)
\end{aligned}$$

These are the fundamental equations for vectorial addition on which most of what follows is based. The values for $s = 2$ may be useful. They are :

$$\begin{aligned}
\zeta_1 &= \kappa_1+\lambda_1 \\
\zeta_2 &= \kappa_2+\lambda_2+2\kappa_1\lambda_1 \\
\zeta_3 &= \kappa_3+\lambda_3+6(\kappa_1\lambda_2+\lambda_1\kappa_2) \\
\zeta_4 &= \kappa_4+\lambda_4+12(\kappa_1\lambda_3+\lambda_1\kappa_3)+30\kappa_2\lambda_2+6(\kappa_1^2\lambda_2+\lambda_1^2\kappa_2)-6\kappa_1^2\lambda_1^2 \\
\zeta_5 &= \kappa_5+\lambda_5+20(\kappa_1\lambda_4+\lambda_1\kappa_4)+90(\kappa_2\lambda_3+\lambda_2\kappa_3)+30(\kappa_1^2\lambda_3+\lambda_1^2\kappa_3)+60(\kappa_1+\lambda_1)\kappa_2\lambda_2- \\
&\qquad -60\kappa_1\lambda_1(\kappa_1\lambda_2+\lambda_1\kappa_2). \quad \ldots \quad (7)
\end{aligned}$$

When $s = 1$ these expressions give the cumulants of the distribution of the variance of the sum of two random variables in terms of the cumulants of the distributions of the two variances.

THE CONDITIONS THAT ADDITION SHOULD BE ASSOCIATIVE

It is clear from the symmetrical character of equations (5) that the addition of random vectors is commutative, that is to say $\mathbf{X}+\mathbf{Y} = \mathbf{Y}+\mathbf{X}$, no matter what are the mean values c_{2r} of $\cos^{2r}\theta$. But unless c_{2r} has the special values given in (2) this addition does not in general obey the associative law, that is to say the distributions of $(\mathbf{X}+\mathbf{Y})+\mathbf{Z}$ and $\mathbf{X}+(\mathbf{Y}+\mathbf{Z})$ are not the same. The first three cumulants are the same, but the fourth is not. To take one example, if μ_r are the cumulants of the distribution of $\mathbf{Z}\,.\,\mathbf{Z}$, then the coefficient of $\kappa_1\lambda_1\mu_2$ in $(\mathbf{X}+\mathbf{Y})+\mathbf{Z}$ is $32\ (6c_2^2+c_4+2c_2c_4)$, and that of $\kappa_2\lambda_1\mu_1$ or $\kappa_1\lambda_2\mu_1$ is $288c_2^2$, from equations (5). The condition for associative addition is that these should be equal, or

$$288c_2^2 = 32(6c_2^2+c_4+2c_2c_4)$$

whence

$$c_4 = \frac{3c_2^2}{1+2c_2}.$$

Thus if $c_2 = s^{-1}$, $c_4 = 3s^{-1}(s+2)^{-1}$.

The same condition can be derived by considering other coefficients. It is, I think, clear that the addition of random vectors whose directions are strictly random in s dimensions is associative. That is to say the addition of random vectors is associative provided equations (2) hold good for the moments of the cosine of the angle between successive vectors. Further (2) is necessary as well as sufficient, for as we consider successive cumulants, each new value of c_{2r} arises as its first power, so it is unambiguously given in terms of the preceding values of c_{2r}.

However there is no need that s should be a positive integer in equations (2). In fact these equations can be satisfied if s is not less than unity. In particular they can be satisfied in two dimensional space. We can postulate that the distribution of the angle ϕ between successive vectors should have the distribution of that between any pair of vectors in random directions in s dimensions, or the distribution whose moments are given by (2) for non-integral s.

THE VECTORIAL CUMULANTS AND THE GENERAL EQUATIONS FOR ASSOCIATIVE ADDITION

We can now proceed to build up general equations from (6). Consider the addition of n random vectors, each with given cumulants; let ζ_r be the r-th cumulant of the distribution of the scalar square of their resultant. If $r > 2$, ζ_r includes some terms of a type not found in (6). For example if to the sum of the two random vectors giving the cumulants of (6) we add a third vector the r-th cumulant of whose scalar square is μ_r, we find :

$$\begin{aligned}\zeta_3 &= \kappa_3+\lambda_3+\mu_3+12s^{-1}[\kappa_1\lambda_2+\lambda_1\kappa_2+(\kappa_1+\lambda_1)\mu_2+\mu_1(\kappa_2+\lambda_2+4s^{-1}\kappa_1\lambda_1)]\\ &= \kappa_3+\lambda_3+\mu_3+12s^{-1}(\kappa_1\lambda_2+\kappa_2\lambda_1+\lambda_1\mu_2+\lambda_2\mu_1+\mu_1\kappa_2+\mu_2\kappa_1)+48s^{-2}\kappa_1\lambda_1\mu_1.\end{aligned}$$

These equations may be written in terms of symmetrical functions or power sums. In what follows $\Sigma\kappa_2\lambda_1$ means the sum of all products of the second cumulant of the distribution of the scalar square of one vector by the first cumulant of that of another, $\Sigma\kappa_1^2\kappa_2$ means the sum of all products of the second cumulant of one such distribution and the square of the first cumulant of the same distribution, and so on. We find:

$$\begin{aligned}\zeta_1 &= \Sigma\kappa_1\\ \zeta_2 &= \Sigma\kappa_2+4s^{-1}\Sigma\kappa_1\lambda_1\\ \zeta_3 &= \Sigma\kappa_3+12s^{-1}\Sigma\kappa_2\lambda_1+48s^{-2}\Sigma\kappa_1\lambda_1\mu_1, \text{ etc.}\end{aligned}$$

$$\begin{aligned}\zeta_1 &= \Sigma\kappa_1\\ \zeta_2 &= \Sigma\kappa_2+2s^{-1}[(\Sigma\kappa_1)^2-\Sigma\kappa_1^2]\\ \zeta_3 &= \Sigma\kappa_3+12s^{-1}[\Sigma\kappa_1\Sigma\kappa_2-\Sigma\kappa_1\kappa_2]+8s^{-2}[(\Sigma\kappa_1)^3-3\Sigma\kappa_1^2\Sigma\kappa_1+2\Sigma\kappa_1^3], \text{ etc.} \qquad \ldots \quad (8)\end{aligned}$$

But these expressions soon become very cumbrous. The expression for ζ_5 in symmetrical functions consists of 12 terms, that in terms of power and product sums consists of 42 terms such as $+960s^{-2}(s+2)^{-1}(s+3)\Sigma\kappa_1(\Sigma\kappa_2)^2$. They are complicated for the same reason as the expression for the 10-th moment of the sum of n random variables with given moments about zero is complicated. These latter complications are of course greatly reduced by the use of cumulants. The question arises whether we can find functions of the cumulants κ_r appertaining to one vector which have suitable additive properties. We can, and they prove to have other properties analogous to those of the cumulants of the distribution of a one-dimensional random variable. I call them the vectorial cumulants θ_r. They can be calculated from (6) without much difficulty by the method of undetermined coefficients given that $\theta_1 = \kappa_1$. We can multiply κ_1 by any constant without affecting the properties of these cumulants, and it might be better to put $\theta_1 = s^{-1}\kappa_1$, but I think the terminology here adopted is preferable. It is easy to show that $\theta_2 = \kappa_2 - 2s^{-1}\kappa_1^2$. Given this, suppose that $\theta_3 = \kappa_3 - x\kappa_1\kappa_2 + y\kappa$, and assume additivity. Then from the third of equations (6)

$$\begin{aligned}&\kappa_3+\lambda_3-x(\kappa_1\kappa_2+\lambda_1\lambda_2)+y(\kappa_1^3+\lambda_1^3) \equiv \zeta_3-x\zeta_1\zeta_2+y\zeta_1^3\\ &\equiv \kappa_3+\lambda_3+12s^{-1}(\kappa_1\lambda_2+\lambda_1\kappa_2)-x(\kappa_1+\lambda_1)(\kappa_2+\lambda_2+4s^{-1}\kappa_1\lambda_1)+y(\kappa_1+\lambda_1)^3.\end{aligned}$$

Hence
$$(\kappa_1\lambda_2+\kappa_2\lambda_1)(12s^{-1}-x)+(\kappa_1^2\lambda_1+\lambda_1^2\kappa_1)(3y-4s^{-1}x) \equiv 0.$$

It follows that $x = 12s^{-1}$, $y = \frac{4}{3}s^{-1}x = 16s^{-2}$.

To determine the 4 coefficients of θ_4 by the same method we have 6 simultaneous linear equations which are consistent. Similarly for the 6 coefficients of θ_5 we have 10 equations. The values of θ_r are:

$$\begin{aligned}\theta_1 &= \kappa_1\\ \theta_2 &= \kappa_2-2s^{-1}\kappa_1^2\\ \theta_3 &= \kappa_3-12s^{-1}\kappa_1\kappa_2+16s^{-2}\kappa_1^3\\ \theta_4 &= \kappa_4-24s^{-1}\kappa_1\kappa_3-24s^{-1}(s+2)^{-1}(s+3)\kappa_2^2+48s^{-2}(s+2)^{-1}(5s+12)\kappa_1^2\kappa_2-\\ &\qquad -48s^{-3}(s+2)^{-1}(5s+12)\kappa_1^4\\ \theta_5 &= \kappa_5-40s^{-1}\kappa_1\kappa_4-120s^{-1}(s+2)^{-1}(s+4)\kappa_2\kappa_3+240s^{-2}(s+2)^{-1}(3s+8)\kappa_1^2\kappa_3+\\ &\quad +1440s^{-2}(s+2)^{-1}(s+4)\kappa_1\kappa_2^2-960s^{-3}(s+2)^{-1}(7s+24)\kappa_1^3\kappa_2+768s^{-4}(s+2)^{-1}\\ &\quad (7s+24)\kappa_1^5\\ &= \kappa_5-40s^{-1}[\kappa_1\kappa_4+3(s+2)^{-1}(s+4)\kappa_2\kappa_3]+240s^{-2}(s+2)^{-1}\kappa_1\\ &\quad [(3s+8)\kappa_1\kappa_3+6(s+4)\kappa_2]-192s^{-3}(s+2)^{-1}(7s+24)\kappa_1^3(5\kappa_2-4s^{-1}\kappa_1^2). \qquad \ldots \quad (9)\end{aligned}$$

The equations for κ_r in .terms of $\theta_1, \theta_2 \ldots, \theta_r$ are :

$$\begin{aligned}\kappa_1 &= \theta_1\\ \kappa_2 &= \theta_2+2s^{-1}\theta_1^2\\ \kappa_3 &= \theta_3+12s^{-1}\theta_1\theta_2+8s^{-2}\theta_1^3\\ \kappa_4 &= \theta_4+24s^{-1}\theta_1\theta_3+24s^{-1}(s+2)^{-1}(s+3)\theta_2^2+144s^{-2}\theta_1^2\theta_2+48s^{-3}\theta_1^4\\ \kappa_5 &= \theta_5+40s^{-1}\theta_1\theta_4+120s^{-1}(s+2)^{-1}(s+4)\theta_2\theta_3+480s^{-2}\theta_1^2\theta_3+\\ &\qquad +960s^{-2}(s+2)^{-1}(s+3)\theta_1\theta_2^2+1920s^{-3}\theta_1^3\theta_2+384s^{-4}\theta_1^5. \qquad \ldots \quad (10)\end{aligned}$$

It will be seen that when $s = 1$, equations (9) agree with equations (4) which were in fact worked out as a check on them, and to enable others to extend them further if desired. Similarly (10) agree with Haldane's (1941) equations (3). They are a good deal simpler than (9).

THE DISTRIBUTION OF THE RESULTANT OF A SAMPLE OF n RANDOM VECTORS, AND THAT OF n UNIT VECTORS

Consider a sample of n random vectors, each drawn from a population in which the r-th cumulant of the distribution of the scalar square is κ_r. To find the cumulants of the distribution of the scalar square of their resultant, we have only to multiply the values of θ_r in (9) by n, and then to apply equations (10). We find :

$$\begin{aligned}
\zeta_1 &= n\kappa_1 \\
\zeta_2 &= 2n(n-1)s^{-1}\kappa_1^2+n\kappa_2 \\
\zeta_3 &= 4n(n-1)s^{-1}\kappa_1[2(n-2)s^{-1}\kappa_1^2+3\kappa_2]+n\kappa_3 \\
\zeta_4 &= 24n(n-1)s^{-1}[2\{(n-2)(n-3)-s(s+2)^{-1}\}s^{-2}\kappa_1^4+2\{3(n-2)+s(s+2)^{-1}\}s^{-1}\kappa_1^2\kappa_2+ \\
&\qquad +\{1+(s+2)^{-1}\}\kappa_2^2+\kappa_1\kappa_3]+n\kappa_4 \\
\zeta_5 &= 8n(n-1)s^{-1}[48(n-2)\{(n-3)(n-4)-5s(s+2)^{-1}\}s^{-3}\kappa_1^5+120\{2(n-2)(n-3)+ \\
&\qquad +(2n-5)s(s+2)^{-1}\}s^{-2}\kappa_1^3\kappa_2+60\{3(n-2)-(n-3)s(s+2)^{-1}\}s^{-1}\kappa_1\kappa_2^2+ \\
&\qquad +30\{2(n-2)+s(s+2)^{-1}\}\kappa_1^2\kappa_3+15\{1+2(s+2)^{-1}\}\kappa_2\kappa_3+5\kappa_1\kappa_4]+n\kappa_5. \qquad \ldots \quad (11)
\end{aligned}$$

Since these equations may be of use in the theory of random migration on a plane surface, I give the values for $s = 2$. They are :

$$\begin{aligned}
\zeta_1 &= n\kappa_1 \\
\zeta_2 &= n(n-1)\kappa_1^2+n\kappa_2 \\
\zeta_3 &= 2n(n-1)\kappa_1[(n-2)\kappa_1^2+3\kappa_2]n\kappa_3 \\
\zeta_4 &= 3n(n-1)[(2n^2-10n+11)\kappa_1^4+2(6n-11)\kappa_1^2\kappa_2+5\kappa_2^2+4\kappa_1\kappa_3]+n\kappa_4 \\
\zeta_5 &= 2n(n-1)[3(n-2)(2n^2-14n+19)\kappa_1^5+30(4n^2-18n+19)\kappa_1^3\kappa_2+15(7n-12)\kappa_1\kappa_2^2+ \\
&\qquad +15(4n-7)\kappa_1^2\kappa_3+45\kappa_2\kappa_3+10\kappa_1\kappa_4]+n\kappa_5. \qquad \ldots \quad (12)
\end{aligned}$$

Finally I give the cumulants of the resultant of n unit vectors with random direction. Here $\kappa_1 = 1$, $\kappa_r = 0 (r > 1)$. They are :

$$\begin{aligned}
\zeta_1 &= n \\
\zeta_2 &= 2n(n-1)s^{-1} \\
\zeta_3 &= 8n(n-1)(n-2)s^{-2} \\
\zeta_4 &= 48n(n-1)[(n-2)(n-3)(s+2)-s]s^{-3}(s+2)^{-1} \\
\zeta_5 &= 384n(n-1)(n-2)[(n-3)(n-4)(s+2)-5s]s^{-4}(s+2)^{-1}. \qquad \ldots \quad (13)
\end{aligned}$$

It will be seen that in (11) ζ_r tends to $2^r(r-1)!\, n^r s^{1-r}\kappa_1^r$ as n tends to infinity, that is to say the distribution approximates to a Gamma, or Pearsonian Type III, distribution.

THE VECTORIAL CUMULANTS AS MEASURES OF DEPARTURE FROM A GAMMA DISTRIBUTION

It can be shown that if $\kappa_r = 2^{r-1}(r-1)!\, s^{1-r}\kappa_1^r$ then all the vectorial cumulants after the first vanish. If the vector **X** has s orthogonal components each normally distributed with mean zero and standard deviation σ, then $\kappa_1 = s\sigma^2$, and $\kappa_r = s2^r(r-1)!$ σ^{2r}, $\sigma^{-2}x$ has a χ^2 distribution with s degrees of freedom. Haldane (1960) gives some values of κ_r for other "parent" distributions and a test for the significance of the departure of κ_2 (which is all that can be estimated with any accuracy in the sample there considered) from zero. If $s = 1$ it is readily seen that $\theta_2 = \mu_4 - 3\mu_2^2$, ($\mu_2$ and μ_4 being moments of x), $\theta_3 = \mu_6 - 15\mu_2\mu_4 + 30\mu_2^3$ and so on, so that θ_r is the $2r$-th cumulant of the distribution of x.

$\theta_r\theta_1^{-r}$ can thus be used as a shape parameter which becomes equal to Fisher's γ_{2r} when $s = 1$. From (9) we see that the sum of n vectors has

$$\theta_1 = \Sigma\kappa_1$$
$$\theta_2 = \Sigma\kappa_2 - 2s^{-1}\Sigma\kappa_1^2$$

etc. Hence $\theta_2\theta_1^{-2}$ tends to zero subject to the conditions of the central limit theorem.

In the case of n unit vectors

$$\theta_1 = n,\ \theta_2 = -2s^{-1}n,\ \theta_3 = +16s^{-2}n,\ \theta_4 = -48s^{-2}(s+2)^{-1}(5s+12)n, \text{ etc.}$$

So $\theta_2\theta_1^{-2} = -2/ns$. The greater the number of dimensions the more rapid the approach to zero. $\theta_r\theta_1^{-r}$ tends to zero with n^{1-r}, but can be very large when r is large.

DISCUSSION

Beckmann (1959) has considered the case of the resultant of a number of unit vectors whose directions in two dimensions have a given distribution. This is complementary to my general case. The present investigation arose from a consideration of Alström's (1958) work on distances between birthplaces of parents and offspring. It is possible however that it may find applications in physics as well as in demography. The velocity of a molecule is a vector, and its scalar square is the proportional to the kinetic energy, so in this case it is quite natural to consider the distribution of the squared velocity, which was of course done by Maxwell. In a perfect gas $s = 3$, and $\theta_r = 0 (r > 1)$. Unfortunately for many purposes we are interested in the small fraction of molecules whose energy exceeds a critical value, and the first few moments of a distribution give little information about them.

I have deliberately kept this paper short, since it is possible that some different symbolism may prove more efficient than my own. However it seems possible that the study of random vectors may throw light on several problems of practical and theoretical statistics. Perhaps the most obvious extension is to consider the resultant of random vectors on the surface of a sphere, and more generally in various kinds of non-Euclidean space.

I have to thank Professor C. R. Rao and Mr. S. R. Varadhan for valuable corrections and suggestions.

REFERENCES

ALSTRÖM, C. H. (1958) : First-Cousin marriages in Sweden 1750–1844 and a study of the population movement in some Swedish sub-populations from the genetic-statistical viewpoint. *Acta Genetica et Statistica Media*, **8**, 295–369.

BECKMANN, P. (1959) : *Acta Technica* (Praha), **4**, 323–335.

HALDANE, J. B. S. (1941) : The cumulants of the distribution of the square of a variate. *Biometrika*, **32**, 199–200.

——— (1960) : Radially symmetrical distributions. (In preparation, to be published in *Sankhyā*).

JEFFREYS, H. (1933) : *The Theory of Probability* (Oxford).

Paper received : February, 1960.

A NOTE ON THE ESTIMATION OF VARIANCE*

By LEO A. GOODMAN
University of Chicago, USA

SUMMARY. A number of different loss functions are considered in the situation where a real valued random variable, X, whose distribution has the invariance property under a change of scale, is observed and where estimators of the form aX, where a is a constant, are to be used to estimate the scale parameter. For each of the loss functions considered, the problem of determining the estimator that minimizes the risk associated with that loss function and/or that is unbiased with respect to that loss function is studied. Some of the results obtained are presented in a sufficiently general form so that they can be used to deal with the problem of estimating the scale parameter of a population in the situation where the form of the population need not be completely known (but certain essential aspects of this form are known) as well as in the situation where the form is completely known. The results obtained are applied to the following special problems : (1) the estimation of the variance of a normal distribution when both the mean and the variance are unknown; (2) the estimation of the variance of a normal distribution when the mean is known; (3) the estimation of the range of a uniform distribution on the interval (o,p) when p is unknown.

1. INTRODUCTION

Consider a random variable X whose distribution depends on an unknown real parameter θ. If the value x of X is observed one makes an estimate, say $f(x)$, and thereby incurs a loss of $W[\theta, f(x)]$ (see, e.g., Wald, 1950); $W[\theta, f(x)]$ is called the "loss function." The risk $R_f(\theta)$ associated with the estimate f is defined to be the expected loss; i.e., $R_f(\theta) = E\{W[\theta, f(x)] | \theta\}$. It is, of course, desirable to choose an estimator f that will minimize the risk. An estimator f is defined as unbiased with respect to the loss function $W[\theta, f(x)]$ if for each θ, $E\{W[\theta', f(x)] | \theta\}$ is minimized when $\theta' = \theta$ (see, e.g., Lehmann, 1951).

Suppose $x_1, x_2, \ldots, x_n$ is a sample of n observations from a normal distribution where both the mean μ and the variance σ^2 are unknown. In the present note, we shall be concerned with, among other things, the problem of estimating σ^2. Lindley (1953) has pointed out that there does not appear to be an estimate of the variance σ^2 that has similar compelling reasons for its adoption as has the arithmetic mean estimate of the mean μ. In other words, the estimate of the variance will depend more on the appropriate loss function.

Let $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $S = \sum_{i=1}^{n} (x_i - \bar{x})^2$. Several estimates have been suggested for σ^2. The maximum likelihood estimate of σ^2 is S/n. The unbiased estimate of σ^2 is $S/(n-1)$. The estimate $S/(n-2)$ has been suggested by Lindley (1953) since it closely approximates the invariant estimate that minimizes the risk when the loss function is $(\log f - \log \sigma^2)^2$ where f is the estimate of σ^2. The present author (1953) has shown that $S/(n+1)$ is the estimate that minimizes the risk (among estimates of the form S/d) when the loss function is $\lambda(\sigma^2)(f-\sigma^2)^2$, where $\lambda(z)$ is any positive

---

*Research carried out at the Statistical Research Center, University of Chicago, under sponsorship of the Statistics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

function of z; i.e., $S/(n+1)$ minimizes the mean square error and also the mean square relative error. The four denominators that have been mentioned, $n+1$, n, $n-1$, $n-2$, have an arithmetic average of $n-1/2$. In the present paper, we shall see that the estimate $S/(n-1/2)$ approximates the estimate f that will minimize the risk when the loss function is $\lambda(\sigma)(\sqrt{f}-\sigma)^2$ and that $\sqrt{S/(n-1/2)}$ approximates the unbiased estimate of σ with respect to the loss function $(\sqrt{f}-\sigma)^2/\sigma^2$.

Lindley (1953) states that the proportional accuracy of σ, the standard deviation, is usually of interest. Although he makes this statement in justification of the loss function $(\log f-\log \sigma^2)^2 = 4(\log\sqrt{f}-\log \sigma)^2$, we might use this same statement to justify the loss function $(\sqrt{f}-\sigma)^2/\sigma^2$. The exact value of the denominator d, for the estimator $f = S/d$, that will minimize the risk when the loss function is $(\sqrt{f}-\sigma)^2/\sigma^2$ is $2\Gamma^2\left(\frac{n+1}{2}\right) \Big/ \Gamma^2\left(\frac{n}{2}\right)$; where $\Gamma(z)$ is the gamma function; numerical values of d are given in Table 1 for $n = 2, 3, \ldots, 10$. This estimator also minimizes the risk when the loss function is of the more general form $\lambda(\sigma)(\sqrt{f}-\sigma)^2$, and the square root of this estimator is an unbiased estimator of σ with respect to the loss function $(\sqrt{f}-\sigma)^2/\sigma^2$. For larger values of n, the value of d is approximately $n-1/2$.

TABLE 1. THE DENOMINATOR THAT MINIMIZES THE RISK WHEN THE LOSS FUNCTION IS $\lambda(\sigma)\,(\sqrt{f}-\sigma)^2$

n	2	3	4	5	6	7	8	9	10
denominator	1.57	2.55	3.53	4.53	5.52	6.52	7.52	8.51	9.51

Table 2 summarizes the various denominators d, for the estimates $f = S/d$, that will be discussed in the present paper. It also gives some optimum properties of each estimate and some approximations to the corresponding values of d. In this table, $\psi(z) = \dfrac{d \log \Gamma(z)}{dz}$.

TABLE 2. THE DENOMINATORS THAT MINIMIZE THE RISK WITH RESPECT TO LOSS FUNCTIONS AND THEIR APPROXIMATIONS

denominator	approximation	loss function for which risk is minimized	loss function with respect to which estimate is unbiased
$2\,\Gamma^2\left(\frac{n+1}{2}\right) / \Gamma^2\left(\frac{n}{2}\right)$	$n-1/2$	$\lambda(\sigma)(\sqrt{f}-\sigma)^2$	$(\sqrt{f}-\sigma)^2/\sigma^2$
$2\exp\left[\psi\left(\frac{n-1}{2}\right)\right]$	$n-2$	$\lambda(\sigma)(\log\sqrt{f}-\log\sigma)^2$	$(\log\sqrt{f}-\log\sigma)^2$
$2\,\Gamma^2\left(\frac{n}{2}\right) / \Gamma^2\left(\frac{n-1}{2}\right)$	$n-3/2$	—	$(\sqrt{f}-\sigma)^2$
$n+1$ $n-1$	$n+1$ $n-1$	$\lambda(\sigma^2)\,(f-\sigma^2)$ —	$(f-\sigma^2)^2/\sigma^4$ $(f-\sigma^2)^2$
$(n-1)\log\left[\frac{1+k}{1-k}\right]/2k$	$(n-1)(1+k^2/3)$	$\begin{cases}0 \text{ when } \lvert f-\sigma^2\rvert \leqslant k\sigma^2\\ 1 \text{ when } \lvert f-\sigma^2\rvert > k\sigma^2\end{cases}$	$\begin{cases}0 \text{ when } \lvert f-\sigma^2\rvert \leqslant k\sigma^2\\ 1 \text{ when } \lvert f-\sigma^2\rvert > k\sigma^2\end{cases}$

We note that the approximate denominators considered range between $n+1$ and $n-2$. For n sufficiently large, this range will be of no great importance from a practical point of view. However, for small values of n, it will be of some importance to choose the denominator that corresponds to the appropriate loss function in a particular situation. So far as we know, the approximations $n-1/2$, $n-3/2$, and $(n-1)(1+k^2/3)$ have not appeared in the statistics literature before now. Neither has there been any reference before this to the unbiased nature of some of the estimates discussed in the present paper.

In the present paper, we shall also study a more general problem of estimation; viz., the estimation of the scale parameter of a population in (a) the situation where the form of the population is given and (b) the situation where its form is not completely given. Some special cases and applications will be discussed in detail.

2. RESULTS

Theorem 1: *Let X be a random variable whose distribution depends on an unknown real parameter θ. Suppose that the expected value $E\{\log X/\theta|\theta\}$ of the natural logarithm of X/θ is a known constant A. Then among all statistics of the form aX, where a is a constant, the risk*

$$R_a(\theta) = E\{\lambda(\theta)(\log aX-\log\theta)^2|\theta\}$$

is uniformly minimized when $a = e^{-A}$.

Proof : We have that

$$R_a(\theta) = E\{\lambda(\theta)[(\log a)^2+2\log a\log(X/\theta)+(\log(X/\theta))^2]\}$$
$$= \lambda(\theta)[(\log a)^2+2A\log a+E\{(\log(X/\theta))^2\}].$$

The risk is a quadratic function of $\log a$ which is minimized when

$$\frac{\partial R_a(\theta)}{\partial a} = \lambda(\theta)[\log a+A]2/a = 0;$$

that is, when $a = e^{-A}$. Q.E.D.

The present author (1953) has proved that, when $E\{X/\theta|\theta\}/E\{(X/\theta)^2|\theta\}$ is a known constant B, then the statistic BX uniformly minimizes the risk $R_a(\theta) = E\{\lambda(\theta)(aX-\theta)^2|\theta\}$; i.e., where the loss function is $\lambda(\theta)(aX-\theta)^2$, rather than $\lambda(\theta)(\log aX-\log\theta)^2$.

Let us now consider the special case where X is a real valued random variable whose distribution has the invariance property under a change of scale; i.e., the probability function of X is $\theta^{-1}f(X/\theta)$, $\theta > 0$, where $f(x)$ is known but the parameter θ which determines the scale of the distribution of X is unknown (see, e.g., Pitman, 1939). The preceding results concerning the loss functions $\lambda(\theta)(\log aX-\log\theta)^2$ and $\lambda(\theta)(aX-\theta)^2$ hold in this case since $E\{\log X/\theta|\theta\}$ and $E\{X/\theta|\theta\}/E\{(X/\theta)^2|\theta\}$ are known. Let us now consider the loss function

$$W_3[\theta, aX] = \begin{cases} 0 & \text{when } |aX-\theta| \leqslant k\theta \\ 1 & \text{when } |aX-\theta| > k\theta \end{cases}$$

where k is a constant. In order to minimize the risk, we must maximize the probability that $|aX-\theta| \leqslant k\theta$. (The reader is referred to Blackwell and Girshick (1954) for a more general treatment of the scale parameter problem). In other words we wish to maximize

$$P = Pr\{|aX-\theta| \leqslant k\theta\} = \int\limits_{(1-k)/a}^{(1+k)/a} f(x)dx.$$

To find the value of the constant a that maximizes P, we study the behaviour of

$$\frac{\partial P}{\partial a} = \{(1-k)f([1-k]/a)-(1+k)f([1+k]/a)\}a^2.$$

In the following section, we shall determine explicitly the value of a that maximizes P for various probability functions $f(x)$.

The reader will note that Theorem 1 does not involve the notion of unbiasedness. The preceding estimates, which minimized the risk with respect to various loss functions, will in general not be unbiased in the usual sense. However, we have the following theorem :

Theorem 2: *Suppose that* $E\{\log X/\theta|\theta\} = A$, *where* A *is known. Then among all statistics of the form* aX, *where* a *is a constant, the only one that is unbiased with respect to the loss function*

$$W_1[\theta, f(x)] = (\log f(x)-\log\theta)^2$$

is $e^{-A}X$ *(which is, by Theorem* 1, *the estimator that uniformly minimizes the risk).*

Proof : We have that

$$\begin{aligned} R_a(\theta') &= E\{[(\log(a/\theta'))^2+2\log(a/\theta')\log X+(\log X)^2]|\theta\} \\ &= (\log(a/\theta'))^2+2\log(a/\theta')[A+\log\theta]+E\{(\log X)^2|\theta\}. \end{aligned}$$

The risk is a quadratic function of $\log\theta'$ which is minimized when

$$\frac{\partial R_a(\theta')}{\partial\log\theta'} = \{-\log a+\log\theta'-[A+\log\theta]\}\,2 = 0;$$

that is, when $\log a + A = \log(\theta'/\theta)$. Hence, we see that $R_a(\theta')$ is minimized when $\theta' = \theta$ if and only if $a = e^{-A}$. Q.E.D.

The present author (1953) has shown that, when $E\{X/\theta|\theta\}/E\{(X/\theta)^2|\theta\} = B$, then BX was the unbiased estimate with respect to the loss function $(f(x)-\theta)^2/\theta^2 = W_2[\theta, f(x)]$.

For the preceding problems, we see, from the results obtained by Lehmann (1951), that the concept of unbiasedness in the usual sense is equivalent to the concept of unbiasedness with respect to the loss function $(f(x)-\theta)^2$. From this fact, or by a simple computation, we see that $\log e^{-A}X = \log X - A$ is an unbiased estimate of $\log \theta$ (see Lindley, 1953).

Let us now return to the case when X is a real valued random variable whose distribution has the invariance property under a change of scale. The preceding results concerning unbiasedness with respect to the loss function $W_1[\theta, f(x)]$ and $W_2[\theta, f(x)]$ hold in this case. Let us now consider the loss function $W_3[\theta, f(x)]$. If our method of determining the appropriate estimate $\left(\text{i.e., by studying the behaviour of } \frac{\partial P}{\partial a}\right)$ leads to one estimate that uniformly minimizes the risk, then we see, from Lehmann's results (1951), that this estimate is also unbiased with respect to $W_3[\theta, f(x)]$.

3. APPLICATIONS

Suppose $x_1, x_2, \ldots, x_n$ is a sample of n observations from a normal distribution where both the mean μ and the variance σ^2 are unknown. Let $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $S = \sum_{i=1}^{n} (x_i - \bar{x})^2$; whence S/σ^2 has a Chi-square distribution with $n-1$ degrees of freedom. Hence,

$$A = E\{\log (S/\sigma^2)|\sigma^2\} = \int_0^\infty (\log x) e^{-x/2} \quad x^{(n-3)/2} dx / 2^{(n-1)/2} \Gamma\left(\frac{n-1}{2}\right)$$

$$= \log 2 + \psi\left(\frac{n-1}{2}\right),$$

where $\Gamma(x)$ is the gamma function and $\psi(x) = \dfrac{d \log \Gamma(x)}{dx}$.

Therefore, the estimate $S/2e^{\psi\left(\frac{n-1}{2}\right)}$ is the invariant estimate that minimizes the risk $R_f(\theta) = E\{\lambda(\theta)(\log f(x) - \log \sigma^2)^2|\sigma^2\}$, and that is unbiased with respect to the loss function $W_1[\sigma^2, f(x)]$. Lindley (1953) states that this estimate may be approximated by $S/(n-2)$ and that the approximation is excellent for $n > 3$; for $n = 2$ and 3 the denominators are 0.28 and 1.12, respectively.

From Goodman (1953) we see that the estimate $S/(n+1)$ is the invariant estimate of σ^2 that minimizes the risk $R_f = E\{\lambda(\sigma^2)(f(x)-\sigma^2)^2|\sigma^2\}$ and that is unbiased with respect to the loss function $W_2[\sigma^2, f(x)]$.

We also see from Goodman (1953) that the estimate

$$f = S2\Gamma^2\left(\frac{n}{2}\right) / (n-1)^2\Gamma^2\left(\frac{n-1}{2}\right)$$

is the invariant estimate of σ^2 that minimizes the risk $R_f = E\{\lambda(\sigma^2)(\sqrt{f(x)}-\sigma)^2|\sigma^2\}$ and that is unbiased with respect to the loss function $W_2[\sigma, \sqrt{f(x)}]$. The denominator d, in this case, is

$$(n-1)^2\Gamma^2\left(\frac{n-1}{2}\right)/2\Gamma^2\left(\frac{n}{2}\right) = 2\Gamma^2\left(\frac{n+1}{2}\right)/\Gamma^2\left(\frac{n}{2}\right)$$

which is approximated by $n-1/2$, as can be seen with the aid of the asymptotic expansion of $\Gamma(x)$. We find that this estimate f may be approximated by $S/(n-1/2)$ and the quality of the approximation is illustrated by the figures in Table 1. (The square root $\sqrt{f}$ of this estimate f is not an unbiased estimate of σ with respect to the loss function $(\sqrt{f}-\sigma)^2$, but it is unbiased with respect to $(\sqrt{f}-\sigma)^2/\sigma^2$; the estimate that is unbiased with respect to $(\sqrt{f}-\sigma)^2$ is the unbiased estimate of σ in the usual sense, viz.,

$$\sqrt{S}\,\Gamma\left(\frac{n-1}{2}\right)/\Gamma\left(\frac{n}{2}\right)\sqrt{2} \text{ (see, Cramér, 1946, p. 383).}$$

Let us now consider the loss function $W_3[\sigma^2, f(x)]$. The invariant estimate that minimizes the risk in this case is

$$S2k/(n-1)\log[(1+k)/(1-k)].$$

Thus, the estimate depends on the value of k, which determines the loss function. When the value of k is small, the denominator of S is approximately $(n-1)(1+k^2/3)$ and, hence, can be approximated by $n-1$. Therefore, when k is small, $S/(n-1)$, which is the unbiased estimate of σ^2, approximates the estimate that minimizes the risk $E\{W_3[\sigma^2, f(x)]|\theta\}$ and that is unbiased with respect to the loss function $W_3[\sigma^2, f(x)]$.

In the case where the population mean μ is known, let $T = \sum_{i=1}^{n} (x_i - \mu)^2$ (see, Hodges and Lehmann, 1951). The preceding considerations would lead to the estimates

$$(1)\quad T/2e^{\psi\left(\frac{n}{2}\right)} \sim T/(n-1)$$

$$(2)\quad T/(n+2)$$

$$(3)\quad T2\Gamma^2\left(\frac{n+1}{2}\right)/n^2\Gamma^2\left(\frac{n}{2}\right) = T\Gamma^2\left(\frac{n+1}{2}\right)/2\Gamma^2\left(\frac{n+2}{2}\right) \sim T/(n+\tfrac{1}{2})$$

$$(4)\quad T2k/n \log\,[(1+k)/(1-k)] \sim T/n \text{ (for small } k\text{)},$$

respectively, for the various loss functions we have considered. (The estimate $\sqrt{T}\,\Gamma\left(\frac{n}{2}\right)/\Gamma\left(\frac{n+1}{2}\right)\sqrt{2}$ is the unbiased estimate of σ in the usual sense).

The following application is of interest in the estimation of production using serial numbers. For some further discussion of this application, the reader is referred to the work by Goodman (1952) and (1954).

Suppose $x_1, x_2, \dots, x_n$ is a sample of n from a uniform distribution on (o, p), where p is unknown. Writing $g = \max(x_1, x_2, \dots, x_n)$, we see that

$$A = E\{\log(g/p)\,|\,p\} = \int_0^1 (\log x)\, nx^{n-1}\,dx = -\int_0^{+\infty} ye^{-y}\,dy/n = -\Gamma(2)/n = -1/n.$$

Therefore, the estimate $ge^{1/n}$ is the invariant estimate that minimizes the risk $R_f(p) = E\{\lambda(p)(\log f(x) - \log p)^2\,|\,p\}$ and that is unbiased with respect to the loss function $W_1[p, f(x)]$. So far as we know, this estimate has not appeared in the statistics literature before now.

From Goodman (1953), we see that the estimate $g(n+2)/(n+1)$ is the invariant estimate of p that minimizes the risk $R_f(p) = E\{\lambda(\theta)(f(x)-p)^2\,|\,p\}$ and that is unbiased with respect to the loss function $W_2[p, f(x)]$.

When the loss function is $W_3[p, f(x)]$, we see that the estimate $g(1+k)$ minimizes the risk and is unbiased (with respect to $W_3[p, f(x)]$). When k is very small, the maximum likelihood estimate g is an approximation to the estimate that minimizes this risk.

REFERENCES

BLACKWELL, D. and GIRSHICK, M. A. (1954): *Theory of Games and Statistical Decisions*, John Wiley and Sons, New York.

CRAMÉR, HARALD (1946): *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

GOODMAN, L. A. (1952): Serial number analysis. *J. Amer. Stat. Ass.*, **47**, 622–634.

——— (1953): A simple method for improving some estimators. *Ann. Math. Stat.*, **22**, 114–117.

——— (1954): Some practical techniques in serial number analysis. *J. Amer. Stat. Ass.*, **49**, 97–112.

HODGES, J. L., JR. and LEHMANN, E.L. (1951): Some applications of the Cramér-Rao inequality. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, California, 13–22.

LEHMANN, E. L. (1951): A general concept of unbiasedness. *Ann. Math. Stat.*, **22**, 587–592.

LINDLEY, D. V. (1953): Statistical inference. *J. Roy. Stat. Soc.* **15**, 30–65.

PITMAN, E. J. G. (1939): The estimation of the location and scale parameter of a continuous population of a given form. *Biometrika*, **30**, 391-421.

WALD, A. (1950) : *Statistical Decision Functions*, John Wiley and Sons, New York.

Paper received : May, 1959.

ON THE ASYMPTOTIC EFFICIENCY OF TESTS AND ESTIMATES

By R. R. BAHADUR

Indian Statistical Institute, Calcutta

SUMMARY. Let $x_1, x_2, \ldots$ be a sequence of independent and identically distributed observations with distribution determined by a real valued parameter θ. For each $n = 1, 2, \ldots$, let $T_n = T_n(x_1, x_2, \ldots, x_n)$ be a statistic such that the sequence $\{T_n\}$ is a consistent estimate of θ. It is shown, under weak regularity conditions on the sample space of a single observation, that the asymptotic effective standard deviation of T_n cannot be less than $[nI(\theta)]^{-\frac{1}{2}}$. The asymptotic effective standard deviation of T_n is defined, roughly speaking, as the solution τ of the equation $P(|T_n-\theta| \geqslant \varepsilon | \theta) = P(|N| \geqslant \varepsilon/\tau)$ when n is large and ε is a small positive number, where N denotes a standard normal variable. It is also shown, under stronger regularity conditions, that the asymptotic effective standard deviation of the maximum likelihood estimate of θ is $[nI(\theta)]^{-\frac{1}{2}}$. These conclusions concerning estimates are derived from certain conclusions concerning the relative efficiency of alternative statistical tests based on large samples.

1. INTRODUCTION

Let X be an abstract sample space of points x, and suppose that the distribution of x is determined by a (not necessarily real valued) parameter θ taking values in a set Θ. Let g be a given real valued function of θ and suppose that it is required to estimate g. For each $n = 1, 2, \ldots$ let $(x_1, x_2, \ldots, x_n) = x_{(n)}$ (say) denote n independent observations on x, and let $X_{(n)}$ denote the sample space of $x_{(n)}$. An *estimate* is defined (without explicit reference to g) to be a sequence $\{T_n\} = \mathbf{T}$ (say), such that $T_n = T_n(x_{(n)})$ is a real valued measurable function on $X_{(n)}$, $(n = 1, 2, \ldots)$. An estimate $\mathbf{T} = \{T_n\}$ is said to be consistent if for each θ and each $\epsilon > 0$, $P(|T_n - g(\theta)| \geqslant \epsilon | \theta) \to 0$ as $n \to \infty$; $\mathbf{T}$ is consistent and asymptotically normal (c.a.n., in short) if for each θ in Θ there exists a sequence $\{\sigma_n\}$ of positive real numbers such that $\lim\limits_{n\to\infty} \sigma_n(\theta) = 0$ and $\lim\limits_{n\to\infty} P([T_n - g(\theta)]/\sigma_n(\theta) \leqslant x | \theta) = P(N \leqslant x)$ for every x, where N denotes a normally distributed random variable with $E(N) = 0$ and $E(N^2) = 1$. $\sigma_n^2(\theta)$ is then called the asymptotic variance of T_n when θ obtains.

The classical theory of estimation from large samples (cf., e.g., Fisher (1922, 1925), Neyman (1949), Gurland and Barankin (1951), LeCam (1953), Kallianpur and Rao (1955); cf. also the references cited by these authors) has been concerned mainly with c.a.n. estimates, the usual criterion of assessment of a particular estimate being its asymptotic variance. As is well known, rigorous theoretical development of the criterion just mentioned has proved full of complications and difficulties. One of the reasons for the difficulties encountered is surely that the asymptotic variance σ_n^2 of a statistic T_n has a very weak (or at least ill-determined) relation to the actual concentration of the distribution of T_n at the value of g —unless , of course, T_n happens to be exactly normally distributed with mean g, and σ_n^2 is the actual variance of T_n.

To put it in another way, in comparing estimates which are c.a.n., but otherwise arbitrary, comparisons of their asymptotic variances appear to lack justification. Indeed, Basu (1956) has given an example where such comparisons seem definitely misleading; this example exhibits two c.a.n. estimates, $\{T_n\}$ and $\{U_n\}$, with asymptotic variances $\{a_n\}$ and $\{b_n\}$ respectively, such that $a_n/b_n \to 0$ as $n \to \infty$ but, for every $\epsilon > 0$,

$$\frac{P(|T_n-g(\theta)| \geqslant \epsilon|\theta)}{P(|U_n-g(\theta)| \geqslant \epsilon|\theta)} \to \infty. \qquad \ldots \quad (1.1)$$

One of the objects of this paper is to suggest an approach, parallel to the classical theory, in which a given c.a.n. estimate $\{T_n\}$ (or indeed any estimate) is discussed in terms of 'effective standard deviations', the latter being scaling constants which are well articulated with the actual probability distributions of the estimating statistics T_n. This approach can be effected in several different ways. The particular version presented here is tentative, and has no claim to logical necessity.

Consider a c.a.n. estimate $\mathbf{T} = \{T_n\}$, and suppose that for a particular n, θ, and $\epsilon > 0$ we wish, for some practical or theoretical reason, to compute $P(|T_n-g| \geqslant \epsilon)$. If σ_n^2 is the asymptotic variance of T_n, $P(|N| \geqslant \epsilon/\sigma_n)$ is an approximation to the probability required. This approximation is, however, of unknown accuracy, and that is precisely why σ_n^2 is in general an unsatisfactory index of the performance of T_n. Suppose we define τ as follows:

Definition 1.1 : For any real valued statistic T_n, any θ, and any $\epsilon > 0$, $\tau = \tau_g(T_n, \epsilon, \theta)$ is the solution of the equation

$$P(|N| \geqslant \epsilon/\tau) = P(|T_n-g(\theta)| \geqslant \epsilon|\theta), \qquad (0 \leqslant \tau \leqslant \infty). \quad \ldots \quad (1.2)$$

Then τ achieves what σ is supposed to do, i.e., the right side of (1.2) can be computed exactly by entering a standard table of the normal distribution with ϵ/τ. Consequently, τ might be called the effective standard deviation of T_n when (a) T_n is regarded as a point estimate of g, (b) θ obtains, and (c) it is required to compute the right side of (1.2). If T_n is exactly normally distributed with mean g then, for every ϵ, $\tau_g(T_n, \epsilon, \theta)$ equals, as it should, the actual standard deviation of T_n.

Although suggested by the study of c.a.n. estimates, Definition 1.1 is applicable to any estimate whatsoever. In particular, if T_n and U_n are any two real valued statistics then $P(|T_n-g(\theta)| \geqslant \epsilon|\theta) \geqslant P(|U_n-g(\theta)| \geqslant \epsilon|\theta)$ if and only if $\tau_g(T_n, \epsilon, \theta) \geqslant \tau_g(U_n, \epsilon, \theta)$. Again, $\mathbf{T} = \{T_n\}$ is a consistent estimate of g if and only if, for each ϵ and each θ in Θ, $\tau_g(T_n, \epsilon, \theta) \to 0$ as $n \to \infty$. In the following development based on Definition 1.1, we consider the class of all consistent estimates. The restriction to c.a.n. estimates is not made henceforth because it is unnecessary, and because the present definitions and conclusions concerning the wider class of estimates may have some bearing on formulations of large sample estimation theory (e.g. Savage (1954)) in which asymptotic normality is not an *a priori* requirement.

Let $\mathbf{T} = \{T_n\}$ and $\mathbf{U} = \{U_n\}$ be two estimates of g. We shall define the upper asymptotic efficiency of $\mathbf{T}$ relative to $\mathbf{U}$ when θ obtains, $\bar{e}_g(\mathbf{T}, \mathbf{U} | \theta)$ say, as follows :

$$\bar{e}_g(\mathbf{T}, \mathbf{U} | \theta) = \overline{\lim_{\varepsilon \to 0}} \ \overline{\lim_{n \to \infty}} \ \{\tau_g^2(U_n, \epsilon, \theta)/\tau_g^2(T_n, \epsilon, \theta)\}. \qquad \ldots \quad (1.3)$$

Similarly, the lower asymptotic relative efficiency, $\underline{e}_g(\mathbf{T}, \mathbf{U} | \theta)$ say, is defined by (1.3) but with $\overline{\lim\limits_{\varepsilon \to 0}} \ \overline{\lim\limits_{n \to \infty}}$ replaced by $\underline{\lim\limits_{\varepsilon \to 0}} \ \underline{\lim\limits_{n \to \infty}}$. In the brackets on the right side of (1.3), and elsewhere in this paper, we take the ratios $0/0$ and ∞/∞, if they occur, to be equal to 1. Then $\underline{e}$ and $\bar{e}$ are always well defined, with $0 \leqslant \underline{e} \leqslant \bar{e} \leqslant \infty$.

The following considerations are relevant to the definitions of the preceding paragraph. In the first place, for given T_n, ϵ, and θ, τ^2 is in many cases of the order $1/n$, so that $\lim\limits_{n \to \infty} \{\tau_g^2(U_n, \epsilon, \theta)/\tau_g^2(T_n, \epsilon, \theta)\}$ exists in such cases, and (roughly speaking) equals the limiting ratio of sample sizes required to obtain comparable probabilities of the event $\{|\text{estimated value—actual value}| \geqslant \epsilon\} = E$ (say). In the second place, when n is large, the distribution of any tolerable estimating statistic will be concentrated in a neighbourhood of g, so that small values of ϵ become of primary interest. Since n is made infinite as the first step in the definition of e, it is therefore appropriate that ϵ then be made to tend to zero.

It may be added here that if we were to let $n \to \infty$ and $\epsilon \to 0$ simultaneously, by setting $\epsilon = \lambda/\sqrt{n}$ where λ is a constant, then the above definition of relative efficiency would reduce to the classical definition for c.a.n. estimates with asymptotic variances proportional to $1/n$. (cf. remark 3 in Section 6).

It is of some interest and importance to the present formulation that if $\{T_n\}$ is any consistent estimate of g then, for large n, τ^2 is approximately inversely proportional to $\log [1/P(E)]$; more precisely, for fixed $\epsilon > 0$ and θ,

$$\frac{2}{\epsilon^2} \log P(|T_n - g(\theta)| \geqslant \epsilon | \theta) = - \frac{1}{\tau_g^2(T_n, \epsilon, \theta)} [1 + o(1)] \qquad \ldots \quad (1.4)$$

as $n \to \infty$, where τ is given by (1.2). This asymptotic relation is an immediate consequence of Lemma 2.3. It follows from (1.4) that definitions and conclusions concerning effective variances τ^2 can be readily phrased in terms of probabilities of deviations instead.

It follows from (1.4), in particular, that if $\mathbf{T} = \{T_n\}$ and $\mathbf{U} = \{U_n\}$ are consistent estimates of g, and $\bar{e}(\mathbf{T}, \mathbf{U} | \theta) < 1$, then, for each sufficiently small $\epsilon > 0$, $P(|T_n - g| \geqslant \epsilon | \theta) > P(|U_n - g| \geqslant \epsilon | \theta)$ for all sufficiently large n; in fact (1.1) holds as $n \to \infty$.

Definition 1.2 : An estimate $\mathbf{T}^* = \{T_n^*\}$ is an asymptotically efficient estimate of g if $\mathbf{T}^*$ is consistent and if, for each θ in Θ, $\bar{e}_g(\mathbf{T}, \mathbf{T}^* | \theta) \leqslant 1$ for all other consistent estimates $\mathbf{T}$.

It is shown in Section 5 that asymptotically efficient estimates exist under fairly general conditions, and that the method of maximum likelihood (m.l.) typically leads to such estimates. The main results may be described as follows. Suppose for simplicity that θ is real valued, that Θ is an interval, and that g is a monotonic and differentiable function with derivative $g'(\theta) \neq 0$. Suppose also that the sample space of a single observation x satisfies certain weak regularity conditions. It is shown that for any consistent $\mathbf{T} = \{T_n\}$ and any θ we then have

$$\underline{\lim}_{\varepsilon \to 0} \; \underline{\lim}_{n \to \infty} \; \{n \, \tau_g^2(T_n, \epsilon, \theta)\} \geqslant [g'(\theta)]^2/I(\theta). \qquad \ldots \quad (1.5)$$

Here $I(\theta)$ corresponds to (and, under certain additional conditions, coincides with) the classical 'information concerning θ contained in x'. It is also shown, under certain additional regularity conditions, that if $\hat{\mathbf{U}}=\{\hat{U}_n\}$ is the m.l. estimate of g then, for each θ,

$$\overline{\lim}_{\varepsilon \to 0} \; \overline{\lim}_{n \to \infty} \; \{n \, \tau_g^2(\hat{U}_n, \epsilon, \theta)\} \leqslant [g'(\theta)]^2/I(\theta). \qquad \ldots \quad (1.6)$$

It follows easily from (1.5) and (1.6) that the m.l. estimate of g is efficient according to Definition 1.2.

Definition 1.3: Let $\mathbf{T} = \{T_n\}$ be an estimate of g, and suppose that there exists a sequence $\{v_n(\theta)\}$ of real numbers ($0 \leqslant v_n \leqslant \infty$ for each n) such that

$$\underline{\lim}_{\varepsilon \to 0} \; \underline{\lim}_{n \to \infty} \; \{\tau_g^2(T_n, \epsilon, \theta)/v_n(\theta)\} = 1 = \overline{\lim}_{\varepsilon \to 0} \; \overline{\lim}_{n \to \infty} \; \{\tau_g^2(T_n, \epsilon, \theta)/v_n(\theta)\}. \qquad \ldots \quad (1.7)$$

Then $v_n(\theta)$ is called the asymptotic effective variance of T_n when θ obtains.

It follows from (1.5), with T_n replaced by $\hat{U}_n$, and (1.6) that $[g'(\theta)]^2/nI(\theta)$ is the asymptotic effective variance of the m.l. estimate of g. In view of (1.4), this conclusion can also be stated as follows : With $\{\hat{U}_n\}$ the m.l. estimate of g,

$$P(|\hat{U}_n - g(\theta)| \geqslant \epsilon \,|\, \theta) = \exp\left(-\tfrac{1}{2}\, n\, \epsilon^2\, I_g\, [1+\delta_n(\epsilon,\theta)]\right) \qquad \ldots \quad (1.8)$$

where

$$I_g(\theta) = I(\theta)/[g'(\theta)]^2 \quad \text{and} \quad \lim_{\varepsilon \to 0} \; \overline{\lim}_{n \to \infty} \; \{|\delta_n(\epsilon, \theta)|\} = 0. \qquad \ldots \quad (1.9)$$

In the following Section 2 some purely analytical lemmas are stated, the main one (Lemma 2.1) being a result of Chernoff (1952) concerning the distribution of a sum of independent and identically distributed random variables. Lemma 2.1, together with the fundamental lemma of Neyman and Pearson, are used in Section 3 to study the asymptotic efficiency of tests of a simple hypothesis against a simple alternative.

It is shown that when the sample size n is large, and the probability of an error of type two is bounded away from 1, the minimum attainable probability of an error of type one is approximately $\exp(-nH)$, where H is one of the information functions introduced by Kullback and Liebler (1951). (cf. also Savage (1954)). An application of this conclusion to the stochastic comparison of tests (Bahadur, 1960b) is given in Section 4. The theorem of Section 3 is shown in Section 5 to lead to the asymptotic inequality (1.5), by using the fact that a consistent estimate of g also provides a consistent test of the value of g. The inequality (1.6) is established, however, by methods typical of estimation theory.

Connections between asymptotic theories of estimation and of testing hypotheses have appeared in the literature from time to time, in more or less concrete forms. The present paper provides an example of a very concrete and explicit connection. Estimation theory appears here as a limiting case of the theory of tests, with the considerations of Section 4 marking the transition from one theory to the other. In particular, the asymptotic efficiency of the m.l. estimate of θ is formally equivalent to certain conclusions to the effect that tests of $\theta = \theta_0$ based on the m.l. estimate are asymptotically efficient against alternatives θ in the neighbourhood of θ_0.

2. LEMMAS

Let $z_1, z_2, \ldots$ be a sequence of independent and identically distributed real valued random variables. For each $n = 1, 2, \ldots$ let

$$S_n = \sum_{i=1}^{n} z_i . \qquad \ldots \quad (2.1)$$

Let $k_1, k_2, \ldots$ be a sequence of constants, and define

$$\alpha_n = P(S_n \geqslant k_n). \qquad \ldots \quad (2.2)$$

Suppose that the following conditions (i)–(iv) are satisfied; these conditions imply, among other things, that $\alpha_n \neq 0$ but $\alpha_n \to 0$ very rapidly as $n \to \infty$. The conditions are (i) that

$$\lim_{n \to \infty} \{k_n/n\} = H \text{ (say), where } -\infty < H < \infty. \qquad \ldots \quad (2.3)$$

Let ϕ denote the moment generating function (m.g.f.) of z, i.e. $\phi(t) = E(e^{tz})$. We suppose (ii) that $\phi(t_1) < \infty$ for some $t_1 > 0$. It then follows from well-known properties of m.g.fs that the set $\{t : \phi(t) < \infty\}$, T say, is an interval which includes positive values. It also follows that ϕ possesses derivatives of all orders in the interior of T, and that the derivatives may be obtained by differentiating $E(e^{tz})$ under the expectation sign. We suppose (iii) that there exists a positive t_0 in the interior of T, such that

$$\frac{\phi'(t_0)}{\phi(t_0)} = H. \qquad \ldots \quad (2.4)$$

Finally, it is assumed (iv) that z is non-degenerate, i.e.

$$P(z = c) < 1 \text{ for every constant } c. \qquad \ldots \quad (2.5)$$

The reader may verify that conditions (ii), (iii) and (iv) are satisfied, in particular, if $\phi(t) < \infty$ for all t, $H > E(z)$ and $P(z > H) > 0$.

Define

$$\psi(t) = e^{-tH}\phi(t) = e^{-tH}E(e^{tz}). \qquad \ldots \quad (2.6)$$

It follows easily from (2.4) and (2.5) that $\psi(t_0) = \rho$ (say) is the minimum value of ψ, and that $0 < \rho < 1$.

The following lemma, due to Chernoff, is basic to this paper.

Lemma 2.1 : *Suppose that*

$$k_n/n = H \text{ for every } n. \qquad \ldots \quad (2.7)$$

Then (a) $\alpha_n \leqslant \rho^n \leqslant [\psi(t)]^n$ *for every n and t, and* (b) *for any given* ϵ, *with* $0 < \epsilon < \rho$, $\alpha_n \geqslant (\rho-\epsilon)^n$ *for all sufficiently large n.*

Proof : For the proof see Chernoff (1952). An alternative proof, and certain refinements of Lemma 2.1, have been given by Bahadur and Ranga Rao (1960).

It follows from Lemma 2.1 that if (2.7) holds then $n^{-1} \log \alpha_n = \log \rho + o(1)$. The following lemma states that this asymptotic formula remains valid when (2.7) is relaxed to (2.3).

Lemma 2.2 : $\lim\limits_{n \to \infty} \{n^{-1} \log \alpha_n\} = \log \rho.$

Proof: It follows from (2.5) that the second derivatives of $\log \phi(t)$ is positive throughout the interior of T, so that $\phi'(t)/\phi(t)$ is strictly increasing and continuous therein. Consequently, by (2.4), the equation $\phi'(t)/\phi(t) = u$ has a positive solution, $f(u)$ say, in the interior of T, provided $|u-H|$ is sufficiently small. Choose and fix a $u > H$ so that $u-H$ is sufficiently small. Then we have $\alpha_n \geqslant P(S_n \geqslant nu)$ for all sufficiently large n, by (2.2) and (2.3). Hence $\underline{\lim}\, n^{-1} \log \alpha_n \geqslant -uf(u)+\log \phi(f(u))$, by an application of Lemma 2.1. Since $f(u)$ is a continuous function of u in a neighbourhood of $u = H$, and since $\phi(t)$ is continuous in t in a neighbourhood of $t = t_0 = f(H)$, it follows by letting $u \to H$ that $\underline{\lim}\, n^{-1} \log \alpha_n \geqslant -Ht_0+\log \phi(t_0) = \log \rho$. By taking $u < H$, a similar argument shows that $\overline{\lim}\, n^{-1} \log \alpha_n \leqslant \log \rho$. Thus $\lim\, n^{-1} \log \alpha_n = \log \rho$, and this completes the proof.

Lemma 2.3: $\log P(|N| \geqslant x) = -\dfrac{x^2}{2}\,[1+o(1)]$ *as* $x \to \infty$.

Proof : This is an immediate consequence of the lemma given in Feller (1957, p. 166).

The next and final lemma is to the effect that the asymptotic effective variance (cf. Section 1) of the sample mean of n independent and identically distributed random variable $z_1, z_2, \ldots$ exists and equals n^{-1} times the actual variance of each z. It is assumed now that the m.g.f. of z exists, i.e. that there exists a $\delta > 0$, such that $\phi(t) = E(e^{tz}) < \infty$ for all t with $|t| < \delta$. We also suppose, without additional loss of generality, that

$$E(z) = 0; \text{ and } 0 < E(z^2) < \infty. \qquad \ldots \quad (2.8)$$

S_n is defined by (2.1) for each n.

Lemma 2.4: *If* $\lambda_n(\epsilon)$ *is defined by*

$$P(|S_n/n| \geqslant \epsilon) = P(|N| \geqslant \epsilon/\lambda_n(\epsilon)) \qquad \ldots \quad (2.9)$$

for $\epsilon > 0$ *and* $n = 1, 2, \ldots$ *then*

$$\lim_{\epsilon\to 0} \lim_{n\to\infty} \{n\, \lambda_n^2(\epsilon)\} = E(z^2). \qquad \ldots \quad (2.10)$$

Equivalently,

$$P(|S_n/n| \geqslant \epsilon) = e^{-\frac{n\varepsilon^2}{2E(z^2)}[1+\delta_n(\varepsilon)]} \qquad \ldots \quad (2.11)$$

where

$$\lim_{\epsilon\to 0} \lim_{n\to\infty} \{\delta_n(\epsilon)\} = 0. \qquad \ldots \quad (2.12)$$

Proof: Consider a fixed $\epsilon > 0$. Let t_1 and t_2 denote the solutions, respectively, of the equations $\phi'(t)/\phi(t) = \epsilon$ and $\phi'(t)/\phi(t) = -\epsilon$. It follows from (2.8) that t_1 and t_2 exist and are uniquely determined for all sufficiently small ϵ. Let $a = \psi(t_1)$ and $b = \psi(-t_2)$, where ψ is defined by (2.6), with H replaced by ϵ therein.

The left side of (2.9) is equal to $P\left(\sum_1^n z_i \geqslant n\epsilon\right) + P\left(\sum_1^n (-z_i) \geqslant n\epsilon\right)$. Applications of Lemma 2.2 show that this equals $a_n^n + b_n^n$, where $a_n \to a$ and $b_n \to b$ as $n \to \infty$. It follows hence that $P(|S_n|/n \geqslant \epsilon) = c_n^n$, where $c_n \to c = \max\{a, b\}$. Consequently,

$$\lim_{n\to\infty} \{n^{-1} \log P(|N| > \epsilon/\lambda_n(\epsilon))\} = \log c, \qquad \ldots \quad (2.13)$$

by (2.9). It follows from (2.9) and (2.13) by Lemma 2.3 that

$$\lim_{n\to\infty} \{n\, \lambda_n^2(\epsilon)\} = \frac{\epsilon^2}{-2\log c}\,. \qquad \ldots \quad (2.14)$$

It follows from the definition of t_1 and (2.8) that $t_1 = \epsilon/\sigma^2 + o(\epsilon)$ as $\epsilon \to 0$, where $\sigma^2 = \phi''(0) = E(z^2)$. Since $\phi(t) = 1 + t^2\sigma^2/2 + o(t^2)$ as $t \to 0$, we have $\log a = -\epsilon t_1 + \log \phi(t_1) = -\epsilon^2/2\sigma^2 + o(\epsilon^2)$. This last expansion is valid also for b, and therefore for c. The desired conclusion (2.10) now follows from (2.14). The equivalence of (2.10) with (2.11), (2.12) is a consequence of Lemma 2.3. This completes the proof.

3. ASYMPTOTIC EFFICIENCY OF TESTS

In this section and the following one we consider the statistical framework described in the first paragraph of Section 1, but no parametric function g is specified. The following three conditions on the sample space X of a single observation x are assumed to hold.

Condition 3.1 : θ is identifiable, i.e. if θ_0 and θ_1 are points in Θ with $\theta_0 \neq \theta_1$ then $P(x \text{ in } A \mid \theta_0) \neq P(x \text{ in } A \mid \theta_1)$ for at least one measurable set $A \subset X$.

Condition 3.2 : The set of alternative distributions of x is dominated, i.e. there exists a σ-finite measure on X, say $\mu(A)$, and a non-negative function $f(x \mid \theta)$ on $X \times \Theta$ such that, for each θ in Θ, f is measurable in x, and such that

$$P(x \text{ in } A \mid \theta = \int_A f(x \mid \theta)\, d\mu \text{ for all } A \subset X. \qquad \dots \quad (3.1)$$

Define
$$z(x \mid \theta_1, \theta_0) = \log\,[f(x \mid \theta_1)/f(x \mid \theta_0)]. \qquad \dots \quad (3.2)$$

As in Section 1 in another context, 0/0 and ∞/∞ are here understood to be equal to 1.

Condition 3.3 : For any two points θ_0 and θ_1 in Θ,

$$E(z^2 \mid \theta_1) < \infty \qquad \dots \quad (3.3)$$

and
$$E(\exp{(\epsilon z)} \mid \theta_1) < \infty \text{ for some } \epsilon = \epsilon(\theta_1, \theta_0) > 0 \qquad \dots \quad (3.4)$$

where z is given by (3.2).

Of these conditions, Condition 3.3 is the only one that is at all restrictive. The condition implies, in particular, that the set of alternative distributions of x is homogeneous, i.e. if $P(A \mid \theta) = 0$ for some θ then $P(A \mid \theta) = 0$ for all θ.

Let θ_0 and θ_1 be any two points of Θ with $\theta_0 \neq \theta_1$. In the remainder of this section we restrict attention to the problem of testing the simple hypothesis that $\theta = \theta_0$ against the simple alternative that $\theta = \theta_1$. For each n, let $x_{(n)}$ denote a sample $(x_1, x_2, \ldots, x_n)$ of n independent observations on x, and let $X_{(n)}$ denote the sample space of $x_{(n)}$, as in Section 1. A *test* is defined to be a sequence $\{W_n\} = \mathbf{W}$ say, such that W_n is a measurable subset of $X_{(n)}$. For any given test $\mathbf{W} = \{W_n\}$, let

$$\alpha_n(\mathbf{W}) \equiv \alpha(W_n) = P(x_{(n)} \text{ in } W_n \mid \theta_0),\ \beta_n(\mathbf{W}) \equiv \beta(W_n) = P(x_{(n)} \text{ not in } W_n \mid \theta_1). \dots \quad (3.5)$$

α_n and β_n are then the probabilities of errors of the first and second kinds, respectively, in using W_n as the critical region.

Definition 3.1 : $\mathcal{C}\ (= \mathcal{C}(\theta_1))$ is the class of all tests **W** such that

$$\overline{\lim_{n\to\infty}}\ \{\beta_n(\mathbf{W})\} < 1. \qquad \dots \quad (3.6)$$

For any test **W** in $\mathcal{C}$,

$$\underline{c}\,(\mathbf{W}) = \underline{\lim_{n\to\infty}}\ \{2n^{-1} \log [1/\alpha_n(\mathbf{W})]\}, \quad \bar{c}(\mathbf{W}) = \overline{\lim_{n\to\infty}}\ \{2n^{-1} \log [1/\alpha_n(\mathbf{W})]\}. \qquad \dots \quad (3.7)$$

$\underline{c}\,[\bar{c}]$ is called the lower [upper] asymptotic slope of the test **W**.

The notion of asymptotic slope has been discussed by Bahadur (1960b) in the special case when the test $\mathbf{W} = \{W_n\}$ is based on a sequence of real valued statistics satisfying certain conditions. This special case is also discussed in the next section of this paper. For the present, let us note that the ratio of the slopes of two tests serves as their relative asymptotic efficiency in the following sense. Let $\mathbf{W}^{(1)} = \{W_n^{(1)}\}$ and $\mathbf{W}^{(2)} = \{W_n^{(2)}\}$ be two tests, and suppose for simplicity that $0 < \underline{c}(\mathbf{W}^{(i)}) = \bar{c}(\mathbf{W}^{(i)}) = c_i$ (say) $< \infty$, $i = 1, 2$. Given ϵ, $0 < \epsilon < 1$, let $M^{(i)}(\epsilon)$ be the least positive integer n such that $\alpha(W_n^{(i)}) \leqslant \epsilon$, where α is defined by (3.5). It then follows easily from the definition (3.7) of c_i that $M^{(i)}(\epsilon) = (-2 \log \epsilon)/(c_i + \delta_i)$, where $\delta_i \to 0$ as $\epsilon \to 0$. Consequently, $\lim_{n\to\infty} \{M^{(1)}(\epsilon)/M^{(2)}(\epsilon)\} = c_2/c_1$, i.e. c_2/c_1 is the limiting ratio of the sample sizes required by $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, respectively, to attain an arbitrarily small probability of an error of the first kind. In particular, $\mathbf{W}^{(1)}$ is more efficient (in the present sense) than $\mathbf{W}^{(2)}$ if $c_1 > c_2$.

The following Theorem 3.1 states in effect that, in the class $\mathcal{C}$, each one of a class of Neyman-Pearson likelihood ratio tests has the maximum slope, and that the numerical value of the maximum slope is $2H$, where

Definition 3.2 : $\qquad H = H(\theta_1, \theta_0) = E[z(x\,|\,\theta_1, \theta_0)\,|\,\theta_1]. \qquad \dots \quad (3.8)$

It follows easily from Conditions 3.1—3.3, the definition (3.2) of z, and the inequality $\log t \leqslant t-1$ for $t \geqslant 0$, that we always have $0 < H < \infty$.

Let $\sigma^2 = \sigma^2(\theta_1, \theta_0)$ denote the variance of $z(x\,|\,\theta_1, \theta_0)$ when θ_1 obtains. It follows from Conditions 3.1 and 3.3 that $0 < \sigma^2 < \infty$. Let a be a constant, $-\infty < a < \infty$, and put $r_n = \exp[nH + \sqrt{n}\, a\, \sigma]$, where H is given by (3.8), $0 < r_n < \infty$. For each n, let $W_n^* = \{x_{(n)} : \prod_1^n f(x_i\,|\,\theta_1) \geqslant r_n \prod_1^n f(x_i\,|\,\theta_0)\}$, and let $\mathbf{W}^* = \{W_n^*\}$.

Theorem 3.1 : (*i*) $\mathbf{W}^*$ *is in* $\mathcal{C}$. (*ii*) $\underline{c}(\mathbf{W}^*) = \bar{c}(\mathbf{W}^*) = 2H$. (*iii*) $\bar{c}(\mathbf{W}) \leqslant 2H$ *for all* **W** *in* $\mathcal{C}$.

Proof : Write $z_i = z(x_i\,|\,\theta_1, \theta_0)$. Then by the definition of W_n^* and (3.5) we have $\beta(W_n^*) = P([\Sigma_1^n z_i - nH]/\sqrt{n}\,\sigma < a\,|\,\theta_1)$. Since H and σ^2 are the mean and variance of z when θ_1 obtains, it follows from the central limit theorem that

$$\lim_{n\to\infty} \beta(W_n^*) = P(N < a). \qquad \dots \quad (3.9)$$

Since a is finite, it follows from (3.9) that $\mathbf{W}^*$ satisfies (3.6), and part (i) is established.

We have $\alpha_n(W_n^*) = P(\Sigma_1^n z_i \geqslant nH+\sqrt{n}\, a\, \sigma | \theta_0)$, by (3.5) and the definition of W_n^*. We shall apply Lemma 2.2 to this α_n. Putting $k_n = nH+\sqrt{n}\, a\, \sigma$, we see that (2.3) is satisfied. Since $E(e^{(1+\epsilon)z}|\theta_0) = E(e^{\epsilon z}|\theta_1)$ by (3.2), it follows from (3.4) that assumption (ii) preceding Lemma 2.1 is satisfied and that $t = 1$ is an interior point of the interval on which $\phi(t) = E(e^{tz}|\theta_0)$ is finite. Since $\phi'(t)/\phi(t) = E(ze^{tz}|\theta_0)/\phi(t)$, it follows from the present definitions of z and H that $\phi'(1)/\phi(1) = H$, so that assumption (iii) is also satisfied, with $t_0 = 1$ (cf.(2.4)). It is readily seen that assumption (iv) must also hold, otherwise Condition 3.1 would be violated. Thus Lemma 2.2 applies. In the present case, $\rho = \psi(t_0) = \exp[-t_0 H].\, \phi(t_0) = \exp[-H]$, so that $n^{-1} \log \alpha_n \rightarrow -H$ as $n \rightarrow \infty$. In view of (3.7), this establishes part (ii).

To prove part (iii), choose and fix a test **W** in $\mathcal{C}$. Then choose and fix constants a and $\epsilon > 0$ such that $\beta_n(\mathbf{W}) \leqslant P(N < a)-\epsilon$ for all sufficiently large n; the existence of a and ϵ is assured by (3.6). Let **W*** be defined as in the paragraph preceding the statement of Theorem 1. It then follows from (3.9) that $\beta(W_n) < \beta(W_n^*)$ for all sufficiently large n. Since W_n^* is a likelihood ratio test, it follows from the lemma of Neyman and Pearson that we must have $\alpha(W_n) > \alpha(W_n^*)$ for all sufficiently large n. Hence $\bar{c}(\mathbf{W}) \leqslant \bar{c}(\mathbf{W}^*)$ by (3.7). Since $\bar{c}(\mathbf{W}^*) = 2H$ by part (ii), this completes the proof of the theorem.

The quantity H has been studied by Kullback and Leibler (1951) in terms of the sample space of a single observation x. They showed, in particular, that if $y = S(x)$ is a statistic defined on X, and if $H_0(\theta_1, \theta_0)$ is the resulting quantity when x is replaced by y in the definition of H, then $H_0 \leqslant H$, with equality if and only if y is a sufficient statistic relative to the two-point parameter space $\{\theta_1, \theta_0\}$. Theorem 3.1 affords the following statistical interpretation of the stated theorem of Kullback and Leibler. Writing $y_i = S(x_i)$, suppose that instead of the original sequence $x_1, x_2, \ldots$ only the sequence $y_1, y_2, \ldots$ is made available to the statistician. Then the maximum asymptotic slope available (i.e. $2H_0$) is the same as before (i.e. $2H$) if S is a sufficient statistic, but is otherwise smaller.

4. THE MAXIMUM SLOPE OF A STANDARD SEQUENCE

In this section we consider the problem of testing the null hypothesis that $\theta = \theta_0$, where θ_0 is a given point of Θ, by means of suitable real valued statistics. Let $T_1, T_2, \ldots$ be a sequence such that T_n is a real valued measurable function of $x_{(n)}$. Write $P(T_n < u | \theta_0) = F_n(u)$.

Definition 4.1 : $\{T_n\}$ is said to be a standard sequence for testing $\theta = \theta_0$ if the following three conditions are satisfied. (I) $\lim\limits_{n \rightarrow \infty} F_n(u) = F(u)$ for every u, where F is a continuous probability distribution function. (II) There exists a function on $(0, \infty)$ into $(0, \infty)$, say f, such that for any sequence $\{u_n\}$ with $u_n \rightarrow \infty$, $u_n^2/n \rightarrow d$, where $0 < d < \infty$, we have $2n^{-1} \log [1-F_n(u_n)] \rightarrow -f(d)$. (III) There exists a function b on Θ such that $b(\theta_0) = 0$ and $0 < b(\theta) < \infty$ for $\theta \neq \theta_0$, and such that $\{T_n/\sqrt{n}\}$

is a consistent estimate of b. If $\{T_n\}$ is a standard sequence, $f(b^2(\theta))$ is called its asymptotic slope when θ obtains, $(\theta \neq \theta_0)$.

Suppose that $\{T_n\}$ is a standard sequence. Then T_n has a limiting distribution if $\theta = \theta_0$, but $T_n \to \infty$ in probability if $\theta \neq \theta_0$, so that large values of T_n are significant when T_n is regarded as a testing statistic. Consequently, given $x_{(n)}$, $1-F_n(T_n(x_{(n)})) = L_n$ (say) is called the level attained by T_n in the given case, i.e. L_n is the probability of obtaining a value of T_n which is greater than or equal to the value actually observed. Write $K_n = -2 \log L_n$. It can be shown that in the null case K_n tends in distribution to a chi-square with two d.f., and that in the non-null case $K_n/n \to f(b^2(\theta))$ in probability.

This last result implies that the ratio of the asymptotic slopes of two alternative standard sequences is, roughly speaking, the inverse ratio of the sample sizes required in order to attain comparable levels of significance in the non-null case. A number of examples of this method of comparison of tests are given by Bahadur (1960a, 1960b). As is stated in these papers, although the method makes no explicit reference to the Neyman-Pearson theory of tests, there is a formal connection with the latter theory. This connection is exploited in the proof of Theorem 4.1 below.

A standard sequence $\{T_n\}$ is said to be non-degenerate at θ_1 if there exists a sequence of constants, $\{k_n\}$ say, such that

$$0 < \varliminf_{n\to\infty} P(T_n < k_n \mid \theta_1) \leqslant \varlimsup_{n\to\infty} P(T_n < k_n \mid \theta_1) < 1. \quad \ldots \quad (4.1)$$

Assuming Conditions 3.1, 3.2, and 3.3, we shall now prove

Theorem 4.1 : *If the standard sequence* $\{T_n\}$ *is non-degenerate at* θ, *its asymptotic slope cannot exceed* $2H(\theta, \theta_0)$ *when* θ *obtains.*

Proof : Choose and fix a $\theta_1 \neq \theta_0$, and suppose $\{T_n\}$ non-degenerate at θ_1. Then there exists a sequence $\{k_n\}$ such that (4.1) holds. For each n, let $W_n = \{x_{(n)} : T_n(x_{(n)}) \geqslant k_n\}$, and regard $\mathbf{W} = \{W_n\}$ as a test of θ_0 against θ_1 (cf. Section 3). Then $\mathbf{W}$ is in class $\mathcal{C}$, by (4.1) and Definition 3.1. We shall show that

$$\underline{c}(\mathbf{W}) = \bar{c}(\mathbf{W}) = f(b^2(\theta_1)). \quad \ldots \quad (4.2)$$

The desired conclusion will then follow from part (iii) of Theorem 3.1.

By condition (III) of Definition 4.1, $T_n/\sqrt{n} \to b(\theta_1)$ in probability, where $0 < b(\theta_1) < \infty$. It follows hence that

$$\lim_{n\to\infty} \{k_n/\sqrt{n}\} = b(\theta_1), \quad \ldots \quad (4.3)$$

for otherwise (4.1) would not hold. Now, $\alpha(W_n) = P(T_n \geqslant k_n \mid \theta_0) = 1-F_n(k_n)$. It follows, therefore, by (4.3) and condition (II) of Definition 4.1, that $2n^{-1} \log [1/\alpha(W_n)] = f(b^2(\theta_1))+o(1)$ as $n \to \infty$, i.e. that (4.2) holds. This completes the proof.

In general, there exists no standard sequence which is optimum in the sense that its slope equals $2H(\theta, \theta_0)$ for each θ in Θ. This is essentially because in general, for given n and given size α, there exists no uniformly most powerful critical region in $X_{(n)}$ for testing $\theta = \theta_0$ against the composite hypothesis $\theta \neq \theta_0$. A special case where such regions do exist, and consequently an optimum standard sequence also exists, is treated by Bahadur (1960a).

It seems fairly certain, however, that under general conditions there exists a standard or nearly standard sequence $\{T^*_n\}$, with slope c^* say, which is locally optimum in the sense that $c^*/2H$ tends to 1 as θ tends to θ_0 through non-null values, i.e. $\{T^*_n\}$ is nearly optimum throughout a neighbourhood of θ_0. This is argued in the subsequent paragraphs of this section. The notion of local optimality seems to be of considerable interest. One reason is that if θ is very distant from θ_0 it is implausible that a large sample will be drawn for testing purposes, so that asymptotic comparisons of alternative standard sequences (whatever such comparisons may be worth) might as well be confined to the immediate vicinity of θ_0. Consequently, a locally optimum sequence is, for practical purposes, an optimum sequence. This argument is, of course, parallel to the argument for letting ϵ tend to zero in the definition (1.3) of the relative efficiency of two estimates of a given parametric function g. Another reason is that (as is already suggested, and as will be clear from the following paragraphs) the problem of local optimality lies well within the relatively unexplored transition stage from the problem of efficient testing to that of efficient estimation.

Suppose that Θ is an open set in the k dimensional Euclidean space of points $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)})$, $1 \leqslant k < \infty$. Suppose that, for fixed $\theta_0 = (\theta_0^{(1)}, \theta_0^{(2)}, \ldots, \theta_0^{(k)})$, $H(\theta, \theta_0)$ is a sufficiently smooth function of θ; to be specific, let D_i denote partial differentiation with respect to $\theta^{(i)}$, and assume that the second order derivative $D_rD_s[H(\theta, \theta_0)]$ exists and is a continuous function of θ, for all $r, s = 1, 2, \ldots, k$. Since $H(\theta_0, \theta_0) = 0$, and $H(\theta, \theta_0) > 0$ for $\theta \neq \theta_0$, it then follows from Taylor's theorem that with $I_{ij}(\theta_0) = \{D_iD_jH(\theta, \theta_0)\}_{\theta=\theta_0}$, and

$$Q(\theta, \theta_0) = \sum_{i,j=1}^{k} I_{ij}(\theta_0)\,(\theta^{(i)} - \theta_0^{(i)})\,(\theta^{(j)} - \theta_0^{(j)}), \qquad \ldots \quad (4.4)$$

we have

$$H(\theta, \theta_0) = \tfrac{1}{2}\,Q(\theta, \theta_0)\,[1 + \epsilon(\theta, \theta_0)] \qquad \ldots \quad (4.5)$$

where $\epsilon \to 0$ as $\theta \to \theta_0$. $\{I_{ij}(\theta_0)\}$ is necessarily a symmetric positive definite matrix.

It follows from the definition (3.8) of H that with $z = z(x\,|\,\theta, \theta_0)$ we have

$$H(\theta, \theta_0) = E(z\,e^z\,|\,\theta_0), \text{ and } E(e^z\,|\,\theta_0) = 1. \qquad \ldots \quad (4.6)$$

Assuming that $D_j[z]$ and $D_iD_j[z]$ exist for each x, and that the operators D_j and D_iD_j commute with the expectation signs in (4.6), $(i, j = 1, 2, \ldots, k)$, it follows from the preceding definition of I_{ij} by a straightforward calculation that

$$I_{ij}(\theta_0) = E(u_i u_j | \theta), \text{ where } u_r = \{D_r[\log f(x|\theta)]\}_{\theta=\theta_0} \quad \ldots \quad (4.7)$$

and $E(u_r | \theta_0) = 0$. Thus $[I_{ij}]$ coincides with the classical information matrix.

Now, for each n, let V_n be a measurable function on $X_{(n)}$ into k dimensional space, say $V_n(x_{(n)}) = (V_n^{(1)}(x_{(n)}), \ldots, V_n^{(k)}(x_{(n)}))$, such that, for each θ in Θ, $\sqrt{n}\,(V_n - \theta)$ tends in distribution to the k-variate normal distribution with mean zero and covariance matrix $\{I^{ij}(\theta)\} = \{I_{ij}(\theta)\}^{-1}$. It is known that these requirements are met if, for example, $V_n =$ the m.l. estimate of θ based on $x_{(n)}$, and if certain regularity conditions (which need not be specified here) are satisfied. Let

$$T_n^*(x_{(n)}) = \sqrt{n \sum_{i,j=1}^{k} I_{ij}(\theta_0)(V_n^{(i)} - \theta_0^{(i)})(V_n^{(j)} - \theta_0^{(j)})}. \quad \ldots \quad (4.8)$$

Then it is easy to see that $\{T_n^*\}$ satisfies condition (I) of Definition 4.1 with $F(u) = P(\chi_k < u)$, where χ_k^2 is a chi-square variable with k d.f., and that (III) is satisfied with $b(\theta) = \sqrt{Q}$, where Q is given by (4.4).

It seems very difficult to verify condition (II) in the general case, and to determine the function f^* associated with $\{T_n^*\}$ by that condition (assuming that such a function exists). Consequently, the exact level attained by T_n^* in a given case, L_n^* say, cannot be treated by the method described in the paragraph following Definition 4.1. However, given $x_{(n)}$, $P(\chi_k \geqslant T_n^*(x_{(n)})) = L_n^0$ say, is an approximation to L_n^* and this approximate level L_n^0 is easy to analyse. It can be shown that Lemma 2.3 remains valid if $|N|$ is replaced by χ_k (cf. Bahadur, 1960b). It follows hence that, with $K_n^0 = -2 \log L_n^0$, K_n^0/n tends in probability to Q. Thus, in a certain approximate sense, Q is the slope of $\{T_n^*\}$. It now follows from (4.5) that, again in a certain approximate sense, $\{T_n^*\}$ is a locally optimum sequence.

In order to show that $\{T_n^*\}$ is exactly a locally optimum sequence, it is necessary to show at least that $-2n^{-1} \log P(T_n \geqslant T_n^*(x_{(n)}) | \theta_0)$ tends in probability to $c^*(\theta)$ (say) when θ obtains, and that $c^*/Q \to 1$ as $\theta \to \theta_0$. It is to be hoped that this can be done by the methods used in the following section, provided $\{T_n^*\}$ is constructed according to (4.8) from the m.l. estimate of θ.

5. THEOREMS ON ESTIMATION

In this section we consider the framework described in the first paragraph of Section 1, with g a given real valued function defined on Θ. It is assumed that Conditions 3.1, 3.2 and 3.3 are satisfied. In the statements of the additional conditions required in this section, θ_0 denotes an arbitrary but fixed point of Θ.

Condition 5.1: Given θ_0, the set

$$\{\theta : \theta \textit{ in } \Theta,\ |g(\theta) - g(\theta_0)| = \epsilon\} = S_g(\epsilon, \theta_0) \text{ (say)} \quad \ldots \quad (5.1)$$

is non-empty for all sufficiently small $\epsilon > 0$.

Definition 5.1: For any $\epsilon > 0$ such that $S_g(\epsilon, \theta_0)$ defined by (5.1) is non-empty,

$$J_g(\epsilon | \theta_0) = \inf \{H(\theta, \theta_0) : \theta \textit{ in } S_g(\epsilon, \theta_0)\}, \quad \ldots \quad (5.2)$$

where H is given by (3.2) and (3.8). For any real $r > 0$,

$$K_g^{(r)}(\theta_0) = \overline{\lim_{\epsilon \to 0}} \left\{ \frac{J_g(\epsilon | \theta_0)}{\epsilon^r} \right\}. \quad \ldots \quad (5.3)$$

Although $K_g^{(r)}$ is always well defined, with $0 \leqslant K_g^{(r)} \leqslant \infty$, the interesting case is when

$$0 < K_g^{(r)}(\theta) < \infty. \quad \ldots \quad (5.4)$$

It is evident from (5.3) that, given θ, (5.4) can hold for at most one value of r. This useful value, if it exists, is usually $r = 2$, and occasionally $r = 1$. It should also be noted here that the useful value of r may vary with θ in a given estimation problem (cf. remark 2 in Section 6).

The following theorem, which gives bounds on the rate of convergence of any consistent estimate, is valid provided only that Conditions 3.1–3.3 and 5.1 are satisfied.

Theorem 5.1: *If* $\mathbf{T} = \{T_n\}$ *is a consistent estimate of* g *then, for every* θ *in* Θ *and every* $r > 0$,

$$\underline{lim}_{\epsilon \to 0} \; \underline{lim}_{n \to \infty} \{(n\, \epsilon^r)^{-1} \log P(|T_n - g(\theta)| \geqslant \epsilon | \theta)\} \geqslant -K_g^{(r)}(\theta). \quad \ldots \quad (5.5)$$

Proof : Choose and fix an r, and a θ_0 in Θ. We shall establish (5.5) at this θ_0. Choose $\epsilon > 0$ so small that $S_g(\epsilon, \theta_0)$ is non-empty, and let θ_1 be a point in the latter set. Let λ be a constant, $0 < \lambda < 1$, and put $\delta = \lambda \epsilon$.

For each n, let $W_n = \{x_{(n)} : |T_n - g(\theta_0)| \geqslant \delta\}$ and regard $\mathbf{W} = \{W_n\}$ as a test of θ_0 against θ_1 (cf. Section 3). Then

$$\beta(W_n) = P(|T_n - g(\theta_0)| < \delta | \theta_1) \quad \ldots \quad (5.6)$$

by (3.5). Now, $|g(\theta_1) - g(\theta_0)| = \epsilon$, so that

$$|T_n - g(\theta_0)| \geqslant |\epsilon - |T_n - g(\theta_1)|| \quad \ldots \quad (5.7)$$

for every $x_{(n)}$. Since $\delta = \lambda\epsilon < \epsilon$, it follows from (5.6), (5.7), and the consistency of T, that $\beta(W_n) \to 0$ as $n \to \infty$. Consequently, W is in the class $\mathcal{C}(\theta_1)$ of Definition 3.1. Hence $\bar{c}(\mathrm{W}) \leqslant 2H(\theta_1, \theta_0)$ by Theorem 3.1. It follows from the definitions of $\bar{c}$ and W that this last conclusion is

$$\underline{\lim}_{n \to \infty} \{n^{-1} \log P(|T_n - g(\theta_0)| \geqslant \delta | \theta_0)\} \geqslant -H(\theta_1, \theta_0). \quad \ldots \quad (5.8)$$

Since θ_1 in $S_g(\epsilon, \theta_0)$ is arbitrary, it follows from (5.8) by (5.2) that

$$\varliminf_{n\to\infty} \{(n\,\delta^r)^{-1} \log P(|T_n - g(\theta_0)| \geqslant \delta|\theta_0)\} \geqslant -\frac{J_g(\epsilon|\theta_0)}{\lambda^r \epsilon^r}. \qquad \dots \quad (5.9)$$

Since ϵ is arbitrary, and $\delta = \lambda\epsilon \to 0$ continuously through positive values as $\epsilon \to 0$, it follows from (5.9) by (5.3) that at θ_0 the left side of (5.5) is not less than $-K_g^{(r)}(\theta_0)/\lambda^r$. If $K_g^{(r)}(\theta_0) = 0$ or ∞, this establishes (5.5) at θ_0; if $0 < K_g^{(r)}(\theta_0) < \infty$, the desired conclusion follows by letting the arbitrary λ tend to 1. This completes the proof.

It is easy to see from the preceding proof that (5.5) is valid not only for consistent estimates but for any $\{T_n\}$ such that $g(\theta_1) \neq g(\theta)$ implies $\varliminf_{n\to\infty} P(|T_n - g(\theta)| \geqslant \delta|\theta_1) < 1$ for all positive $\delta < |g(\theta_1) - g(\theta)|$.

A different but equally effortless generalisation of Theorem 5.1 is to the case when g is not necessarily real valued. Suppose that g is a function on Θ into a metric space, and that Condition 5.1 is satisfied, with $|g(\theta_1) - g(\theta_0)|$ = the distance between $g(\theta_1)$ and $g(\theta_0)$. With the obvious definitions of a consistent estimate of g, and of $K_g^{(r)}(\theta)$, the preceding proof of (5.5) applies verbatim.

In the remainder of this section it will be necessary to assume that θ takes values in a Euclidean space, and that $H(\theta, \theta_0)$ and $g(\theta)$ are sufficiently smooth functions of θ. For simplicity, we shall consider only the case when θ is a real valued parameter. The parallel development when θ is a k-dimensional vector is discussed briefly in remark 4 of Section 6.

Condition 5.2: θ is real valued. Θ is an open interval. For θ in the neighbourhood of θ_0,

$$H(\theta, \theta_0) = \frac{(\theta - \theta_0)^2}{2} I(\theta_0) + o((\theta - \theta_0)^2), \qquad \dots \quad (5.10)$$

where $0 < I(\theta_0) < \infty$. g is a differentiable function of θ.

It follows by taking $k = 1$ in the paragraphs containing (4.4)–(4.7) that, under certain regularity conditions, I defined by (5.10) coincides with the classical 'information concerning θ contained in x', but these additional conditions are not required at present.

Theorem 5.2: *If* $\mathbf{T} = \{T_n\}$ *is a consistent estimate of* g, *then* (1.5) *holds for each* θ *in* Θ.

Proof: Choose and fix a θ_0 in Θ. We shall establish the inequality (1.5) at θ_0. Since $0 < I(\theta_0) < \infty$, we may assume $g'(\theta_0) \neq 0$, for otherwise the inequality is trivial. Since $g'(\theta_0) \neq 0$, it is plain that Condition 5.1 is satisfied. Moreover, for each sufficiently small $\epsilon > 0$, we can choose a θ_ϵ in $S_g(\epsilon, \theta_0)$ so that $\theta_\epsilon \to \theta_0$ as $\epsilon \to 0$. Since $J_g(\epsilon|\theta_0) \leqslant H(\theta_\epsilon, \theta_0)$ for every ϵ by (5.2), we have

$$\frac{J_g(\epsilon|\theta_0)}{\epsilon^2} \leqslant \frac{H(\theta_\epsilon, \theta_0)}{[g(\theta_\epsilon) - g(\theta_0)]^2}, \qquad \dots \quad (5.11)$$

It follows from (5.10) by letting $\epsilon \to 0$ in (5.11) that

$$K_g^{(2)}(\theta_0) \leqslant \tfrac{1}{2}I(\theta_0) \,/\, [g'(\theta_0)]^2. \qquad \ldots \quad (5.12)$$

It follows readily from (1.4), (5.5) with $r = 2$, and (5.12) that (1.5) holds at θ_0, and this completes the proof.

Theorem 5.2 states in effect that the asymptotic effective variance (Definition 1.3) of a consistent estimate of g cannot be less than $[g']^2/nI$. The remainder of this section is devoted to showing that, under certain additional conditions, the asymptotic effective variance of the m.l. estimate is equal to this lower bound. The conclusion to be established is, in a sense, stronger and more precise than the conclusion that the m.l. estimate of g is a c.a.n. estimate with asymptotic variance $[g']^2/nI$. The regularity conditions which we shall require are essentially a combination of Wald's conditions for the consistency of the m.l. estimate of θ (Wald, 1949), of Cramér's conditions for the asymptotic normality of a root of the likelihood equation (Cramér, 1946) and of conditions required in order to apply the lemmas of Section 2 in certain ways.

Let $L(\theta\,|\,x) = \log f(x\,|\,\theta)$, where f is given by (3.1), and let $D \equiv \partial/\partial\theta$.

Condition 5.3 : (i) For each x in X, $L(\theta\,|\,x)$ is a thrice differentiable function of θ. (ii) With

$$u(x\,|\,\theta_0) = \{DL\}_{\theta=\theta_0} \text{ and } v(x\,|\,\theta_0) = \{D^2L\}_{\theta=\theta_0}, \qquad \ldots \quad (5.13)$$

we have

$$E(u\,|\,\theta_0) = 0,\ E(u^2\,|\,\theta_0) = -\,E(v\,|\,\theta_0) = I(\theta_0). \qquad \ldots \quad (5.14)$$

(iii) There exists a measurable function $w(x\,|\,\theta_0)$, and a neighbourhood of θ_0, such that

$$|\,D^3L\,| \leqslant w(x\,|\,\theta_0) \qquad \ldots \quad (5.15)$$

for every x in X and every θ in the neighbourhood. (iv) The m.g.fs of u, v, and w exist when θ_0 obtains.

In Condition 5.3 (iv) and in Condition 5.4 below, the finite existence of m.g.fs is required only in a neighbourhood of the origin.

In the following, for any real valued function h defined on Θ and any set $\Omega \subset \Theta$, we write $h(\Omega) = \sup\,\{h(\theta) : \theta \text{ } in \text{ } \Omega\}$. In particular, with $z(x\,|\,\theta,\,\theta_0)$ defined by (3.2),

$$z(x\,|\,\Omega,\,\theta_0) = \sup\,\{z(x\,|\,\theta,\,\theta_0) : \theta \text{ } in \text{ } \Omega\}. \qquad \ldots \quad (5.16)$$

Since $z(x\,|\,\theta,\,\theta_0)$ is measurable in x for each θ and, according to Condition 5.3(i), continuous in θ for each x, it is easy to see that $z(x\,|\,\Omega,\,\theta_0)$ is measurable (but not necessarily finite valued).

Condition 5.4 : The m.g.f. of $z(x\,|\,\Theta,\,\theta_0)$ exists when θ_0 obtains.

Since $0 \leqslant z(x\,|\,\Theta,\,\theta_0) \leqslant \infty$, this condition is equivalent to $E(e^{tz}\,|\,\theta_0) < \infty$ for some $t > 0$. The reader may verify that if X is a discrete sample space, and $\sum_x [P(\{x\}\,|\,\theta_0)]^p < \infty$ for some power $p < 1$, then Condition 5.4 is satisfied.

In order to state the next and final regularity condition, suppose $\Theta = \{\theta : a < \theta < b\}$, where $-\infty \leqslant a < b \leqslant \infty$.

Condition 5.5 : For each x in X, $f(x|\theta)$ tends to a limit as θ tends to a or to b. With

$$f(x|a) = \lim_{\theta \to a} \{f(x|\theta)\}, \quad f(x|b) = \lim_{\theta \to b} \{f(x|\theta)\} \qquad \ldots \quad (5.17)$$

we have $\mu\{x : f(x|a) \neq f(x|\theta_0)\} > 0$ and $\mu\{x : f(x|b) \neq f(x|\theta_0)\} > 0$.

Now, for any $n = 1, 2, \ldots$ and any $x_{(n)} = (x_1, x_2, \ldots, x_n)$ in $X_{(n)}$, let $L_n(\theta|x_{(n)})$ denote the logarithm of the likelihood function given $x_{(n)}$, i.e.

$$L_n(\theta|x_{(n)}) = \log \prod_{i=1}^{n} f(x_i|\theta) = L(\theta|x_1) + \ldots + L(\theta|x_n). \qquad \ldots \quad (5.18)$$

Let $M_n(x_{(n)})$ be the (possibly empty) set of points θ in Θ such that $L_n(\theta|x_{(n)}) = L_n(\Theta|x_{(n)})$.

Definition 5.2: A sequence $\hat{\mathbf{T}} = \{\hat{T}_n\}$ is said to be an m.l. estimate of θ if, for each n, $\hat{T}_n$ is a measurable function on $X_{(n)}$ into Θ, and if $\hat{T}_n(x_{(n)})$ is in $M_n(x_{(n)})$ whenever the latter set is non-empty. A sequence $\hat{\mathbf{U}} = \{\hat{U}_n\}$ is said to be an m.l. estimate of $g(\theta)$ if $\hat{U}_n \equiv g(\hat{T}_n)$, where $\hat{T}_n$ is an m.l. estimate of θ.

We can now state

Lemma 5.1: *There exists an m.l. estimate of* θ.

Given n and $x_{(n)}$, we can surely choose a point $\hat{T}_n$ in Θ, in such a way that $\hat{T}_n$ is in M_n if the latter set is non-empty. Lemma 5.1 asserts that this choice can be exercised (for those $x_{(n)}$ for which $M_n(x_{(n)})$ is empty or consists of more than one point) so that the resulting function of $x_{(n)}$ is measurable. The proof consists in giving a constructive definition of $\hat{T}_n$ based on the continuity of the likelihood function and the separability of Θ. This proof is omitted because it is rather long and uninteresting, and because the arguments of the following paragraphs apply to any given m.l. estimate. (cf. LeCam, 1956).

In what follows, $\{\hat{T}_n\}$ is a given m.l. estimate of θ. It is assumed that θ_0 obtains, where θ_0 is an arbitrary but fixed point of Θ.

Lemma 5.2: *Given* $h > 0$, *there exists* ρ, $0 < \rho < 1$, *such that*

$$P(|\hat{T}_n - \theta_0| \geqslant h) < \rho^n \qquad \ldots \quad (5.19)$$

for all sufficiently large n.

This lemma implies the strong consistency of the m.l. estimate of θ. Its proof is along the lines of Wald's proof of consistency (Wald, 1949).

Proof : Consider a θ in Θ with $\theta \neq \theta_0$. We shall show that

$$E(e^{tz(x\,|\,\theta,\,\theta_0)}) < 1 \quad \text{for } 0 < t < 1. \qquad \ldots \quad (5.20)$$

Letting $\phi(t)$ denote the expectation in (5.20), we have $\phi(0) = \phi(1)$, by the definition of z and the present assumption that θ_0 obtains. Since ϕ is an m.g.f., ϕ is convex in [0, 1]. Suppose that $\phi(t_1) = 1$ for some t_1 in (0, 1). It then follows from convexity that $\phi = 1$ for all t in (0, 1) so that $\phi''(t) = E(z^2e^{tz}) = 0$ in that interval. Hence $P(z = 0) = 1$, i.e. $P(f(x\,|\,\theta) = f(x\,|\,\theta_0)\,|\,\theta_0) = 1$. Consequently, $\mu\{f(x\,|\,\theta) \neq f(x\,|\,\theta_0)\} = 0$, and this is contrary to Condition 3.1.

Next, let $f(x\,|\,a)$ be defined by (5.17) and $z(x\,|\,a,\,\theta_0)$ as usual by (3.2), $-\infty \leqslant z \leqslant \infty$. We shall show that (5.20) continues to hold with θ replaced by a. Let $\theta_1,\ \theta_2,\ \ldots$ be a sequence in Θ which tends to a. Writing ϕ_i for ϕ when θ is replaced by θ_i in the preceding paragraph, we have $\phi_i(1) = 1$ for each i. Writing $z = z(x\,|\,a,\,\theta_0)$, it follows hence by (5.17) and Fatou's lemma that with $\phi(t) \equiv E(e^{tz})$ we have $\phi(1) \leqslant 1$. Let $\chi = 1$ if $z > -\infty$ and $\chi = 0$ otherwise, and let $\phi_0(t) = E(\chi e^{tz})$. Then $\phi_0(t) \leqslant \phi(t)$ for all t, with equality for $t > 0$. Suppose that there exists a t_1 with $0 < t_1 < 1$ such that $\phi(t_1) \geqslant 1$. We then have $\phi_0(t_1) \geqslant 1$, $\phi_0(0) \leqslant 1$, and $\phi_0(1) \leqslant 1$. Since ϕ_0 is convex on [0, 1], it follows that $\phi_0 = 1$ on this interval. Now $\phi_0(0) = 1$ implies $E(\chi) = 1$, i.e. $P(z > -\infty) = 1$. Consequently, $\phi(t) = \phi_0(t) = 1$ on [0, 1]. It follows hence, as in the preceding paragraph, that $P(z = 0) = 1$, and hence $\mu\{f(x\,|\,a) \neq f(x\,|\,\theta_0)\} = 0$, which contradicts the second part of Condition 5.5. The argument of this paragraph obviously remains valid when a is replaced by b.

It is therefore established that (5.20) holds for each $\theta \neq \theta_0$ in $\Theta^* = \{\theta : a \leqslant \theta \leqslant b\}$, with $z(x\,|\,a,\,\theta_0)$ and $z(x\,|\,b,\,\theta_0)$ defined as in the preceding paragraph. Now choose and fix a t such that $0 < t < 1$ and such that $E(\exp[tz(x\,|\,\Theta,\,\theta_0)]) < \infty$; such a t exists by Condition 5.4. It then follows from (5.20), by an application of Lebesgue's dominated convergence theorem, that for each θ in Θ^* with $\theta \neq \theta_0$ there exists an interval containing θ, $\mathscr{J}(\theta)$ say, such that with

$$z^*(x\,|\,\theta) = \sup\,\{z(x\,|\,\theta_1,\,\theta_0) : \theta_1 \text{ in } \mathscr{J}(\theta)\}, \quad \rho(\theta) = E(e^{tz^*}) \qquad \ldots \quad (5.21)$$

we have $\rho(\theta) < 1$, and such that $\mathscr{J}(\theta)$ is open in Θ^*.

Given $h > 0$, let Ω be the set $\{\theta : \theta \text{ in } \Theta^*,\ |\theta - \theta_0| \geqslant h\}$. If Ω is empty, (or indeed if $\Omega - \{a\} - \{b\}$ is empty) then $|\hat{T}_n - \theta_0| \geqslant h$ is an impossible event, so that (5.19) holds for every n with $\rho = \frac{1}{2}$ (say). Suppose now that Ω is non-empty. Since Ω is a compact subset of Θ^*, and Ω does not contain θ_0, it is possible to find a finite number of points in Ω, say $\theta_1,\ \theta_2,\ \ldots,\ \theta_k$ such that $\Omega \subset \mathscr{J}(\theta_1) + \mathscr{J}(\theta_2) + \ldots + \mathscr{J}(\theta_k)$, where $\mathscr{J}(\theta)$ is defined for each $\theta \neq \theta_0$ in the preceding paragraph.

Now consider a fixed n and $x_{(n)}$, and suppose first that there do exist θ values in Θ for which $L_n(\theta) = L_n(\Theta)$, i.e. M_n is non-empty. Suppose that $|\hat{T}_n - \theta_0| \geqslant h$. Then $\hat{T}_n$ is in M_n and also in Ω. Consequently, $L_n(\Omega) \geqslant L_n(\hat{T}_n) = L_n(\Theta) \geqslant L_n(\theta_0)$.

Thus

$$L_n(\Omega\,|\,x_{(n)}) \geqslant L_n(\theta_0\,|\,x_{(n)}). \qquad \dots \quad (5.22)$$

Suppose next that M_n is empty, i.e. $L_n(\hat{T}_n) \neq L_n(\Theta)$. In that case, $L_n(\Theta) = \max\,\{L_n(a), L_n(b)\} > L_n(\theta_0)$; hence (5.22) continues to hold, since a and b are included in Ω. Thus $|\hat{T}_n - \theta_0| \geqslant h$ always implies (5.22).

Since $\Omega \subset \mathcal{J}(\theta_1) + \dots + \mathcal{J}(\theta_k)$, (5.22) implies

$$\max_{1\leqslant j\leqslant k} \{L_n(\mathcal{J}(\theta_j)\,|\,x_{(n)})\} \geqslant L_n(\theta_0\,|\,x_{(n)}). \qquad \dots \quad (5.23)$$

Now, for any θ, $L_n(\mathcal{J}(\theta)\,|\,x_{(n)}) - L_n(\theta_0\,|\,x_{(n)})$ is less than or equal to $\sum_{i=1}^{n} z^*(x_i\,|\,\theta)$, by (5.18) and (5.21). Consequently, (5.23) implies

$$\max_{1\leqslant j\leqslant k} \left\{ \sum_{i=1}^{n} z^*(x_i\,|\,\theta_j) \right\} \geqslant 0. \qquad \dots \quad (5.24)$$

Let $A_n^{(j)}$ denote the event that $\sum_{i=1}^{n} z^*(x_i\,|\,\theta_j) \geqslant 0$. Then (5.24) is equivalent to $A_n^{(1)} + A_n^{(2)} + \dots + A_n^{(k)}$. Consequently, by the preceding paragraphs,

$$P(|\hat{T}_n - \theta_0| \geqslant h) \leqslant \sum_{j=1}^{k} P(A_n^{(j)}). \qquad \dots \quad (5.25)$$

It follows from the definition of $A_n^{(j)}$ by taking $H = 0$ and $z = z^*(x\,|\,\theta_j)$ in Lemma 2.1 that $P(A_n^{(j)}) \leqslant [\rho(\theta_j)]^n$, where $\rho(\theta_j)$ is given by (5.21). Let $\rho_0 = \max\,\{\rho(\theta_1), \dots, \rho(\theta_k)\}$. Then $\rho_0 < 1$, and the right side of (5.25) does not exceed $k\,\rho_0^n$. Choose a ρ such that $\rho_0 < \rho < 1$. Then $k\rho_0^n < \rho^n$ for all sufficiently large n. Consequently, by (5.25), (5.19) holds for all sufficiently large n. This establishes Lemma 5.2.

It might appear at first sight that a refinement of the preceding argument would lead to the best possible ρ in (5.19) and that by determining the dependence of this ρ on h it would be possible to determine the asymptotic effective variance of $\hat{T}_n$, but this is not the case. The reason is essentially that the proof of Lemma 5.2 reduces Θ in effect to a finite set. When Θ is finite, the m.l. estimate of θ has necessarily a discrete distribution, and statistics with discrete distributions are liable to have large effective standard deviations. It is therefore necessary to take into account the fact that Θ is a continuum. This is done in the proof of the following lemma by using Condition 5.3 and exhibiting the m.l. estimate as a root of the likelihood equation.

Lemma 5.3: *Given* δ, $0 < \delta < I(\theta_0)$, *there exists a* $\rho < 1$ *such that, for every* $\epsilon > 0$,

$$P(|\hat{T}_n - \theta_0| \quad \epsilon) \leqslant P(|\zeta_n| \geqslant \epsilon[I(\theta_0) - \delta]\,) + \rho^n \qquad \ldots \quad (5.26)$$

for all sufficiently large n, *where*

$$\zeta_n = \frac{\sum_{i=1}^{n} u(x_i|\theta_0)}{n}. \qquad \ldots \quad (5.27)$$

Proof : Write

$$\eta_n = I(\theta_0) + \frac{\sum_{i=1}^{n} v(x_i|\theta_0)}{n}, \quad \xi_n = \frac{\sum_{i=1}^{n} w(x_i|\theta_0)}{n} \qquad \ldots \quad (5.28)$$

where v and w are given by (5.13) and (5.15). Choose and fix an $h > 0$ such that Θ contains the interval $\{\theta : |\theta - \theta_0| < h\}$, (5.15) holds for every θ in this last interval, and $\delta/h > E(w|\theta_0)$. Since v and w possess m.g.fs when θ_0 obtains (Condition 5.3), and $E(v) = -I(\theta_0)$ by (5.14), it follows from (5.28) by means of Lemma 2.1 that

$$P(\eta_n \geqslant \frac{\delta}{2}) \leqslant \rho_1^n,\ P(\eta_n \leqslant -\frac{\delta}{2}) \leqslant \rho_2^n,\ P(\xi_n > \frac{\delta}{h}) \leqslant \rho_3^n \qquad \ldots \quad (5.29)$$

for every n, where each ρ_i is less than 1.

Suppose that for given n and $x_{(n)}$ we have

$$|\hat{T}_n - \theta_0| < h,\ L_n(\hat{T}_n) = L_n(\Theta)\,. \qquad \ldots \quad (5.30)$$

Write $\Delta_n(\theta) = D[L_n(\theta)]$. Since $\hat{T}_n$ is in Θ, and Θ is open, it follows from (5.30) by Taylor's theorem that there exists a θ^* with $|\theta^* - \theta_0| < |\hat{T}_n - \theta_0| < h$ such that $\Delta_n(\theta_0) + (\hat{T}_n - \theta_0)\,\Delta_n'(\theta_0) + \frac{1}{2}(\hat{T}_n - \theta_0)^2\,\Delta_n''(\theta^*) = \Delta_n(\hat{T}_n) = 0$. It follows hence by referring to (5.13), the definition of Δ_n, and (5.27), (5.28), that

$$(\hat{T}_n - \theta_0)[I(\theta_0) + r_n] = \zeta_n \qquad \ldots \quad (5.31)$$

where

$$|r_n| \leqslant |\eta_n| + \tfrac{1}{2} h\, \xi_n. \qquad \ldots \quad (5.32)$$

It is not required here that r_n be a measurable function of $x_{(n)}$, or that (5.30) be a measurable event in $X_{(n)}$.

Let A_n denote the event $|\hat{T}_n - \theta_0| \geqslant h$, B_n the event $L_n(\hat{T}_n) \neq L_n(\Theta)$, and C_n the event $|\eta_n| + \frac{1}{2} h\, \xi_n \geqslant \delta$. As is shown in the proof of Lemma 5.2, each of the events A_n and B_n implies (5.22), and the probability of (5.22) is $\leqslant \rho_0^n$ for all sufficiently large n, where $\rho_0 < 1$. C_n implies at least one of the three events whose probabilities are considered in (5.29). We conclude that there exists a $\rho < 1$ and a measurable event E_n such that $A_n + B_n + C_n$ implies E_n, and such that $P(E_n) \leqslant \rho^n$ for all sufficiently large n.

For any given ϵ, the left side of (5.26) does not exceed $P(x_{(n)}$ *not in* $E_n, |\hat{T}_n - \theta_0| \geqslant \epsilon) + P(E_n)$. It follows hence from (5.31), (5.32), and the preceding paragraph that (5.26) holds for all sufficiently large n, and this completes the proof.

We can now establish

Theorem 5.3: *If* $\{\hat{U}_n\}$ *is an m.l. estimate of* g, *then* (1.6) *holds whenever* $g'(\theta) \neq 0$.

Proof : Choose and fix a δ, $0 < \delta < I(\theta_0)$, and write $a = I(\theta_0) - \delta$. It follows from (5.27) by Lemma 2.4 and Condition 5.3 that, given $\rho < 1$, $n^{-1} \log P(|\zeta_n| \geqslant \epsilon a)$ tends to a limit $> \log \rho$ as $n \to \infty$, provided only that $\epsilon > 0$ is sufficiently small. It follows hence from (5.26) that, with $A_{n,\epsilon} = \{x_{(n)} : |\hat{T}_n - \theta_0| \geqslant \epsilon\}$, we have

$$\overline{\lim_{n \to \infty}} \{n^{-1} \log P(A_{n,\epsilon})\} \leqslant \lim_{n \to \infty} \{n^{-1} \log P(|\zeta_n| \geqslant \epsilon a)\} \qquad \dots \quad (5.33)$$

for all sufficiently small ϵ. Multiply (5.33) with $1/\epsilon^2$, and let $\epsilon \to 0$. It then follows from (5.14) and (5.27) by Lemma 2.4 that

$$\overline{\lim_{\epsilon \to 0}} \quad \overline{\lim_{n \to \infty}} \{(n \epsilon^2)^{-1} \log P(A_{n,\epsilon})\} \leqslant -\tfrac{1}{2} \frac{a^2}{I(\theta_0)}. \qquad \dots \quad (5.34)$$

Since $a = I(\theta_0) - \delta$, and δ is arbitrary, it follows from (5.34) that

$$\overline{\lim_{\epsilon \to 0}} \ \overline{\lim_{n \to \infty}} \{(n \epsilon^2)^{-1} \log P(A_{n,\epsilon})\} \leqslant -\tfrac{1}{2} I(\theta_0). \qquad \dots \quad (5.35)$$

In view of (1.4), (5.35) is equivalent to (1.6) at θ_0 in the special case $g(\theta) \equiv \theta$. To treat the general case, suppose $\{\hat{U}_n\}$ is an m.l. estimate of g. Then $\hat{U}_n \equiv g(\hat{T}_n)$, where $\{\hat{T}_n\}$ is an m.l. estimate of θ. Suppose that $g(\theta_0) \neq 0$. Choose and fix a $\lambda > 1$. It is easy to see that, for all sufficiently small $\epsilon > 0$, $|g(\theta) - g(\theta_0)| \geqslant \epsilon$ implies $|\theta - \theta_0| \geqslant \delta$, where $\delta = \epsilon / \lambda |g'(\theta_0)|$. Let $B_{n,\epsilon} = \{x_{(n)} : |\hat{U}_n - g(\theta_0)| \geqslant \epsilon\}$. Then $B_{n,\epsilon}$ implies $A_{n\delta}$ so that $P(B_{n,\epsilon}) \leqslant P(A_{n,\delta})$. Since $\delta/\epsilon \equiv 1/\lambda |g'|$, it follows from (5.35) that

$$\overline{\lim_{\epsilon \to 0}} \ \overline{\lim_{n \to \infty}} \{(n \epsilon^2)^{-1} \log P(B_{n,\epsilon})\} \leqslant -\tfrac{1}{2} \frac{I(\theta_0)}{\lambda^2 [g'(\theta_0)]^2}. \qquad \dots \quad (5.36)$$

Since $\lambda > 1$ is arbitrary, (5.36) implies

$$\overline{\lim_{\epsilon \to 0}} \ \overline{\lim_{n \to \infty}} \{(n \epsilon^2)^{-1} \log P(B_{n,\epsilon})\} \leqslant -\tfrac{1}{2} \frac{I(\theta_0)}{[g'(\theta_0)]^2}. \qquad \dots \quad (5.37)$$

It follows from (5.37), the definition of $B_{n,\epsilon}$, and (1.4) that (1.6) holds at θ_0, and this completes the proof of Theorem 5.3.

6. FURTHER REMARKS ON ESTIMATION

The following remarks concern the definitions and conclusions of Sections 1 and 5.

1. It is clear from (1.4), and from the proofs of Section 5, that the present study of estimation is based in effect on the logarithm of the probability of a deviation of ϵ or more from the value being estimated. However, when this last probability is the criterion, the interposition of the logarithm means that only the dominant features of the estimates being compared are taken into account. In particular, there may exist two alternative estimates $\mathbf{T} = \{T_n\}$ and $\mathbf{U} = \{U_n\}$, such that both $\mathbf{T}$ and $\mathbf{U}$ are efficient estimates of g according to Definition 1.2, but such that $\mathbf{T}$ is actually more efficient in the sense that

$$\overline{\lim_{\varepsilon \to 0}} \; \overline{\lim_{n \to \infty}} \left\{ \frac{P(|T_n - g(\theta)| \geqslant \epsilon | \theta)}{P(|U_n - g(\theta)| \geqslant \epsilon | \theta)} \right\} < 1 \qquad \ldots \quad (6.1)$$

for all θ. This is the case if, for example, x is real valued and normally distributed with mean θ and variance 1, $g(\theta) = \theta$, $T_n = (x_1 + \ldots + x_n)/n = \bar{x}_n$, and $U_n = \bar{x}_{m_n}$, where $m_1, m_2, \ldots$ is a sequence of positive integers such that $m_n \leqslant n$ for every n, $\lim_{n \to \infty} \{m_n/n\} = 1$, and $\lim_{n \to \infty} \{n - m_n\} = \infty$.

It is possible to formulate asymptotic relative efficiencies in terms of limits such as the left side of (6.1), but the analysis then required seems very difficult, even in the simplest cases.

2. The following example shows that, in general, the r for which (5.4) holds depends on θ.

Suppose that x is real valued, $X = \{x : 0 < x < 1\}$, $\Theta = \{\theta : 0 < \theta < 1\}$, μ = Lebesgue measure on X (cf. Condition 3.2), and

$$f(x|\theta) = \begin{cases} \dfrac{1}{2\theta} & \text{if } 0 < x \leqslant \theta \\[2ex] \dfrac{1}{2(1-\theta)} & \text{if } \theta < x < 1. \end{cases} \qquad \ldots \quad (6.2)$$

Let $g(\theta) \equiv \theta$. It is readily seen that Conditions 3.1, 3.2, 3.3 and 5.1 are satisfied, so that Theorem 5.1 applies.

Let us write K instead of K_g, since here g is the identity function. A straightforward but lengthy calculation, which is omitted, shows that

$$K^{(2)}(\theta) = \begin{cases} 2 & \text{if } \theta = \frac{1}{2} \\ \infty & \text{if } \theta \neq \frac{1}{2}, \end{cases} \qquad \ldots \quad (6.3)$$

and that

$$K^{(1)}(\theta) = \begin{cases} 0 & \text{if } \theta = \tfrac{1}{2} \\ \dfrac{1}{2\delta}\left[\dfrac{(2\delta-1)}{(1-\delta)} + \log\left(\dfrac{1-\delta}{\delta}\right)\right] & \text{if } \theta \neq \tfrac{1}{2}, \end{cases} \qquad \ldots \quad (6.4)$$

where $\delta = \max\,\{\theta, 1-\theta\}$.

It is plain from (6.2) that θ is the median of the distribution of x. Consequently, if $T_n^* =$ the median of the sample values $\{x_1, x_2, \ldots, x_n\}$, $\mathbf{T}^* = \{T_n^*\}$ is a consistent estimate of θ. By expressing the distribution of T_n^* in terms of the binomial distribution, and applying Lemma 2.2 to the latter distribution, it can be shown that

$$\lim_{\epsilon \to 0}\ \lim_{n \to \infty}\ \{(n\,\epsilon^2)^{-1} \log P(|T_n^* - \theta| \geqslant \epsilon\,|\,\theta)\} = -\frac{1}{2\delta^2} \qquad \ldots \quad (6.5)$$

for every θ. It follows from (6.3) and (6.5) that $\mathbf{T}^*$ is efficient when $\theta = \frac{1}{2}$. It does *not* follow from (6.4) and (6.5) alone that $\mathbf{T}^*$ is inefficient for other values of θ. This is because Theorem 5.1 only provides a bound which may or may not be attainable in a given case.

It would be interesting to know whether, in the present case, there exists an estimate which is efficient according to Definition 1.2, and if so, whether efficiency can be established by means of Theorem 5.1.

3. Let $\mathbf{T} = \{T_n\}$ be a c.a.n. estimate of g, and suppose that, when θ obtains, the asymptotic variance of T_n is $v(\theta)/n$, where $0 < v < \infty$. It was generally believed, following the work of R. A. Fisher, that we must then have $v(\theta) \geqslant [g'(\theta)]^2/I(\theta)$ for all θ. This belief was shown to be erroneous by J. L. Hodges, who constructed examples of estimates which were superefficient in the sense that $v(\theta) \leqslant [g'(\theta)]^2/I(\theta)$ for all θ, and

$$v(\theta) < [g'(\theta)]^2/I(\theta) \qquad \ldots \quad (6.6)$$

for some values of θ. A detailed study of superefficiency is given by LeCam (1953), who showed that in general the set of values θ for which (6.6) holds must be a set of Lebesgue measure zero.

It follows from Theorem 5.2 that the phenomenon of superefficiency depends on a certain lack of uniformity in the approach to normality. To see this, let λ be a positive constant. Then

$$\lim_{n \to \infty} P(\sqrt{n}\,|T_n - g(\theta)| \geqslant \lambda\,|\,\theta) = P(|N| \geqslant \lambda/\sqrt{v(\theta)}) \qquad \ldots \quad (6.7)$$

by the asymptotic normality of $\mathbf{T}$. It follows from (6.7) and Definition 1.1 that

$$\lim_{n \to \infty} \left\{ n\ \tau_g^2(T_n, \frac{\lambda}{\sqrt{n}}, \theta)\right\} = v(\theta) \qquad \ldots \quad (6.8)$$

for every λ and θ. Suppose, for a given θ, that (6.8) holds uniformly in λ. It then follows from (1.5) that (6.6) cannot hold for that θ.

4. Suppose that Θ is an open set of the k dimensional Euclidean space of points $\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)})$, and that the conditions stated in the paragraph

containing (4.4)—(4.5) are satisfied. Suppose that the given real valued function $g(\theta)$ possesses continuous partial derivatives $\partial g/\partial\theta^{(i)} = h_i(\theta)$ (say) for $i = 1, 2, \ldots, k$.

The conditions just stated generalise Condition 5.2. If Conditions 3.1—3.3 are also satisfied,

$$\underline{\lim}_{\varepsilon \to 0} \; \underline{\lim}_{n \to \infty} \; \{n\,\tau_g^2(T_n, \epsilon, \theta)^2\} \geqslant \sum_{i,j=1}^{k} I^{ij}(\theta)\, h_i(\theta)\, h_j(\theta) \qquad \ldots \quad (6.9)$$

for any consistent estimate of g. The proof of (6.9) is exactly the same as that of Theorem 5.2, with θ_ε in the proof of Theorem 5.2 so chosen that the vector $\theta_\varepsilon - \theta_0$ is proportional to the vector $(r_1, r_2, \ldots, r_k)$ which minimises $\sum\limits_{i,j=1}^{k} I_{ij}(\theta_0)\, r_i r_j$ subject to the condition $\sum\limits_{i=1}^{k} h_i(\theta_0)\, r_i = 1$ (say).

It is a little more difficult, but quite possible, to generalise Theorem 5.3, i.e., to show (under certain conditions) that if $\{\hat{U}_n\}$ is an m.l. estimate of g then

$$\overline{\lim}_{\varepsilon \to 0} \; \overline{\lim}_{n \to \infty} \; \{n\, \tau_g^2(\hat{U}_n, \epsilon, \theta)\} \leqslant \sum_{i,j=1}^{k} I^{ij}(\theta)\, h_i(\theta)\, h_j(\theta) \qquad \ldots \quad (6.10)$$

whenever the right side of (6.10) is positive. The main difficulty is in formulating a satisfactory generalisation of Condition 5.5; once this is done the proof of (6.10) proceeds along the same lines as that of Theorem 5.3.

REFERENCES

BAHADUR, R. R. (1960a) : Simultaneous comparison of the optimum and sign tests of a normal mean. *Hotelling Festschrift*, Stanford University Press, (Stanford).

——— (1960b) : Stochastic comparison of tests. *Ann. Math. Stat.*, **31** pp.—.

BAHADUR, R. R. and RANGA RAO, R. (1960) : On deviations of the sample mean. *Ann. Math. Stat.*, **31**, pp.—.

BARANKIN, E. W. and GURLAND, J. (1951) : On asymptotically normal efficient estimates. University of California Publications in Statistics. Vol. 1, No. 6, 89–130.

BASU, D. (1956) : On the concept of asymptotic efficiency. *Sankhyā*, **17**, 193.

CHERNOFF, H. (1952) : A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, **23**, 493–507.

CRAMÉR, H. (1946) : *Mathematical Methods of Statistics*, Princeton University Press.

FELLER, W. (1957) : *An Introduction to Probability Theory and its Applications*, Vol. 1, 2nd ed., John Wiley and Sons, (New York).

FISHER, R. A. (1922) : On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.*, London, Series A, **222**, 309–368.

——— (1925) : Theory of statistical estimation. *Proc. Cambridge Phil. Soc.*, **22**, 700–725.

KALLIANPUR, G. B. and RAO, C. R. (1955) : On Fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā*, **15**, 331.

KULLBACK, S. and LEIBLER, R. A. (1951) : On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

LECAM, L. (1953) : On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. University of California Publications in Statistics. Vol. 1, No. 11, 277–330.

——— (1956) : On the asymptotic theory of estimation and testing hypotheses. Proc. Third Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.

NEYMAN, J. (1949) : Contribution to the theory of the χ^2 test. Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, 239–273.

SAVAGE, L. J. (1954) : *The Foundations of Statistics*, John Wiley and Sons, (New York).

WALD, A. (1949) : Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.*, **20**, 595-601.

Paper received : February, 1960.

ON BOUNDED INFINITELY DIVISIBLE RANDOM VARIABLES

By GLEN BAXTER
University of Minnesota, USA
and
J. M. SHAPIRO
The Ohio State University, USA

SUMMARY. Necessary and sufficient conditions for the existence of upper or lower bounds for an infinitely divisible random variable are found. The results are applied to give conditions for the monotonicity of the sample functions of certain stochastic processes.

1. Introduction

Let X be an infinitely divisible random variable and let $\varphi(t)$ be its characteristic function. By the formula of Levy and Khintchine

$$\varphi(t) = \exp\left\{i\gamma t+\int_{-\infty}^{\infty}\left(e^{itu}-1-\frac{itu}{1+u^2}\right)\frac{1+u^2}{u^2}\,dG(u)\right\} \qquad \dots \quad (1.1)$$

where γ is a real constant and $G(u)$ is a bounded non-decreasing function. In this paper we find conditions on $G(u)$ which are necessary and sufficient for the existence of an upper or lower bound for X. As an important special case we show that if $G(+\infty)-G(-0) > 0$, then $P\{X > A\} > 0$ for all real numbers A, i.e. X is not bounded from above. The well known result that if X is a bounded infinitely divisible random variable, then X is constant with probability one, will follow from our theorems. The results are applied to give information about the monotonicity of the sample functions of certain stochastic processes.

2. Preliminary notation and the main results

Formula (1.1) can be written in the following form (see Gnedenko and Kolmogorov, 1954).

$$\log\varphi(t) = i\gamma(\tau)t-\sigma^2t^2/2+\int_{-\infty}^{-\tau}(e^{itu}-1)dM(u)+\int_{\tau}^{\infty}(e^{itu}-1)dN(u)+$$

$$+\int_{-\tau}^{0-}(e^{itu}-1-itu)\,dM(u)+\int_{0+}^{\tau}(e^{itu}-1-itu)dN(u), \qquad \dots \quad (2.1)$$

where

$$\left.\begin{aligned} M(u) &= \int_{-\infty}^{u} \frac{1+z^2}{z^2}\, dG(z) \qquad \text{for } u<0 \\ N(u) &= -\int_{u}^{\infty} \frac{1+z^2}{z^2}\, dG(z) \qquad \text{for } u>0 \\ \sigma^2 &= G(+0)-G(-0) \\ \gamma(\tau) &= \gamma + \int_{|u|<\tau} u\,dG(u) - \int_{|u|\geqslant\tau} 1/u \quad dG(u) \end{aligned}\right\} \qquad \ldots \quad (2.2)$$

and where $\pm\tau$ are continuity points, respectively, of $N(u)$ and $M(u)$. In the remainder of the paper X will always denote an infinitely divisible random variable with associated $G(u)$, $M(u)$, $N(u)$, σ^2 and $\gamma(\tau)$ given by (1.1), (2.1) and (2.2). We now state our main result.

Theorem 1 : *In order that X be bounded from above, i.e. there exists a constant A such that $P\{X>A\}=0$, it is necessary and sufficient that*

$$(i) \quad N(u) = 0 \text{ for all } u>0$$

$$(ii) \quad \sigma^2 = 0$$

$$(iii) \quad \lim_{\varepsilon\to 0^+} \int_{-1}^{-\varepsilon} M(u)du < \infty.$$

The corresponding theorem for bounds from below is :

Theorem 2 : *In order that X be bounded from below, it is necessary and sufficient that*

$$(i) \quad M(u) = 0 \text{ for all } u<0$$

$$(ii) \quad \sigma^2 = 0$$

$$(iii) \quad \lim_{\varepsilon\to 0^+} \int_{\varepsilon}^{1} N(u)du > -\infty.$$

The proofs of Theorems 1 and 2 will be given in the next section.

We now state three corollaries of particular interest.

Corollary 1 : *If $G(+\infty) > G(-0)$, then for any real number A, $P\{X>A\}>0$.*

Corollary 2 : *If $P\{X>A\}>0$ for every real number A, then $G(+\infty) > G(-\epsilon)$ for every $\epsilon>0$.*

Corollaries 1 and 2 have obvious counterparts for random variables X that are unbounded from below. The third corollary was referred to in the introduction. (For a proof by a different method see Pakshirajan and Chatterjee (1956)).

Corollary 3 : *If X is a bounded infinitely divisible random variable, then X is constant with probability one.*

The proofs of these corollaries are given in Section 4.

3. Proofs of theorems 1 and 2

The proof of Theorem 1 is contained in four lemmas. The first three of these show the necessity of the conditions (*i*)—(*iii*) while the fourth shows the sufficiency. The proof is based on a well known theorem (Theorem 3 below) from the theory of limits of sums of independent random variables. (Gnedenko and Kolmogorov, 1954, p. 124). We state this theorem after introducing the following notation :

Let (X_{nk}) $k = 1, 2, \ldots, k_n$; $n = 1, 2, \ldots$ be a double sequence of infinitesimal (i.e. $\lim_{n\to\infty} \max_{1\leqslant k\leqslant k_n} P\{|X_{nk}|>\epsilon\} = 0$ for every $\epsilon > 0$) random variables with distribution functions $F_{nk}(x)$ such that for each $n, X_{n1}, \ldots, X_{nk_n}$ are independent. Let $S_n = X_{n1}+\ldots+X_{nk_n}$ and let $F_n(x)$ be the distribution function of S_n. Let $F(x)$ be the distribution function of an infinitely divisible random variable X with associated $G(u)$, $M(u)$, $N(u)$, σ^2 and $\gamma(\tau)$.

Theorem 3 : *In order that* $\lim_{n\to\infty} F_n(x) = F(x)$ *at every continuity point of* $F(x)$, *it is necessary and sufficient that*

(1) *at continuity points of* $M(x)$ *and* $N(x)$

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} F_{nk}(x) = M(x), \qquad x < 0$$

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} (F_{nk}(x)-1) = N(x), \qquad x > 0$$

$$(2) \quad \lim_{\varepsilon\to 0} \overline{\lim_{n\to\infty}} \sum_{k=1}^{k_n} \left\{ \int_{|x|<\varepsilon} x^2 dF_{nk}(x) - \left(\int_{|x|<\varepsilon} x dF_{nk}(x) \right)^2 \right\} = \sigma^2$$

$$(3) \quad \lim_{n\to\infty} \sum_{k=1}^{k_n} \int_{|x|<\tau} x dF_{nk}(x) = \gamma(\tau).$$

We now proceed to state and prove the lemmas.

Lemma 1 : *If there exists a continuity point* $x_0 > 0$ *of* $N(x)$ *for which* $N(x_0) < 0$, *then* $P\{X > A\} > 0$ *for all real numbers* A.

Proof : Since X is infinitely divisible we can find a double sequence of random variables (X_{nk}) $k = 1, 2, \ldots, n$; $n = 1, 2, \ldots$ such that for each n, $X_{n1}, \ldots, X_{nn}$ are independent, identically distributed and such that $S_n = X_{n1}+\cdots+X_{nn}$ has the same distribution as X. It follows that the (X_{nk}) are infinitesimal (see Gnedenko and Kolmogorov, 1954, p. 58). Thus by Theorem 3

$$\lim_{n\to\infty} n(1-F_{n1}(x_0)) = -N(x_0) > 0. \qquad \ldots \quad (3.1)$$

For n sufficiently large (3.1) implies that $P\{x_{n1} > 0\} \geqslant P\{X_{n1} > x_0\} = 1-F_{n1}(x_0) > 0$, so that $P\{X > 0\} = P\{S_n > 0\} \geqslant [P\{X_{n1} > 0\}]n > 0$. Thus, $P\{X > 0\} > 0$ for any infinitely divisible random variable with the given $N(x)$. For any real number A the random variable $X-A$ is infinitely divisible with the same $N(x)$ associated with it as with X. Thus $P\{X-A > 0\} = P\{X > A\} > 0$, and the proof of Lemma 1 is completed.

Lemma 2 : *If* $\sigma^2 > 0$, *then for every real number* A, $P\{X > A\} > 0$.

Proof : By (1.1) we may write

$$\varphi(t) = e^{-\sigma^2 t^2/2} \exp \left\{ i\gamma t + \int_{-\infty}^{\infty} \left(e^{itx} - 1 - \frac{itx}{1+x^2} \right) \frac{1+x^2}{x^2} d\tilde{G}(x) \right\}$$

where

$$\tilde{G}(x) = \begin{cases} G(x) & x < 0 \\ G(x) - \sigma^2 & x \geqslant 0 \end{cases}$$

and where $\sigma^2 > 0$. Thus the random variable X may be considered as a sum of two independent random variables X_1 and X_2 where X_1 is Gaussian with mean zero and variance $\sigma^2 > 0$. Now surely there exists a $B > 0$ such that $P\{X_2 \geqslant -B\} > 0$. But since X_1 is Gaussian, for any A we have $P\{X_1 > A+B\} > 0$. Thus, $P\{X > A\} = P\{X_1 + X_2 > A\} \geqslant P\{X_1 > A+B,\ X \geqslant -B\} = P\{X_1 > A+B\}.\ P\{X_2 \geqslant -B\} > 0$. This completes the proof of Lemma 2.

Lemma 3 : *If*

$$\lim_{\varepsilon \to 0} \int_{-1}^{-\varepsilon} M(u) du = \infty \quad \dots \quad (3.2)$$

then for every real number A, $P\{X > A\} < 0$.

Proof : (By contradiction) suppose that there exists an A_0 such that $P\{X > A_0\} = 0$. As in the proof of Lemma 1 by a consideration of the random variable $X - A_0$ we may assume without loss of generality that $A_0 = 0$. Consider the same double sequence of random variables (X_{nk}) as in the proof of Lemma 1. Now since $P\{X > 0\} = 0 = P\{S_n > 0\} \geqslant [P\{X_{n1} > 0\}]^n$, it follows that $P\{X_{n1} > 0\} = 0$. By Theorem 3, at continuity points of $M(u)$

$$\lim_{n \to \infty} nF_{n1}(u) = M(u) \qquad u < 0$$

$$\lim_{n \to \infty} n \int_{|x| < \tau} x\, dF_{n1}(x) = \gamma(\tau) \quad \dots \quad (3.3)$$

and since $F_{n1}(x) = 1$ for $x > 0$ we have from an integration parts

$$n \int_{|x|<\tau} x \, dF_{n1}(x) = n \int_{-\tau}^{0} x \, dF_{n1}(x) = n\tau F_{n1}(-\tau) - n \int_{-\tau}^{0} F_{n1}(x)dx. \qquad \ldots \ (3.4)$$

Using (3.3), (3.4), and the fact that $-\tau$ is a continuity point of $M(u)$ we deduce that there exists a finite limit

$$\lim_{n\to\infty} n \int_{-\tau}^{0} F_{n1}(x)dx. \qquad \ldots \ (3.5)$$

Now $nF_{n1}(x) \leqslant nF_{n1}(-\epsilon)$ for all $x \leqslant -\epsilon$ so that by bounded convergence

$$\lim_{n\to\infty} n \int_{-\tau}^{-\varepsilon} F_{n1}(x)dx = \int_{-\tau}^{-\varepsilon} M(x)dx \leqslant \lim_{n\to\infty} n \int_{-\tau}^{0} F_{n1}(x)dx. \qquad \ldots \ (3.6)$$

Taking the limit as $\epsilon \to 0$ on both sides of (3.6) we find that

$$\lim_{\varepsilon\to 0} \int_{-\tau}^{-\varepsilon} M(x)dx \leqslant \lim_{n\to\infty} n \int_{-\tau}^{0} F_{n1}(x)dx < \infty.$$

This contradicts the hypothesis of the lemma. Hence, $P\{X > A\} > 0$ for every real number A, and the lemma is proved.

Taken together Lemmas 1–3 show that if any one of the conditions of Theorem 1 is not satisfied, then $P\{X > A\} > 0$ for every real number A, i.e. X is not bounded from above.

We now turn to the sufficiency part of Theorem 1.*

Lemma 4 : *If conditions (i)–(iii) of Theorem 1 are satisfied, then* $P\{X > A_0\} = 0$ *where the real number* A_0 *is given by*

$$A_0 = \gamma + \int_{-\infty}^{0} \frac{|u|}{1+u^2} dM(u).$$

Proof : Let X be an infinitely divisible random variable with associated $N(u) = 0$ for $u > 0$, $\sigma^2 = 0$; $M(-0) < \infty$, and γ given by (2.1). Let

$$\lambda = \int_{-\infty}^{0} \frac{|u|}{1+u^2} dM(u).$$

---

*The authors wish to thank the referee for suggesting this simplified proof as well as that of the sufficiency of Theorem 4.

It follows that (cf. Gnedenko and Kolmogorov (1954) p. 84) the characteristic function, $\varphi(t)$ of Y can be written in the form :

$$\varphi(t) = \exp\left\{ \int_{-\infty}^{0-} (e^{itu}-1)dM(u)+i(\lambda+\gamma)t \right\}.$$

It follows that the distribution function of $Y-(\lambda+\gamma)$ is given by

$$F(x) = e^{-M(-0)} \sum_{k=0}^{\infty} \frac{M^{(k)}(x)}{k\,!}$$

where $M^{(k)}(x)$ is the k-th iterated convolution of $M(x)$. Since $M(x) \equiv M(-0)$ for $x \geqslant 0$ we see that

$$P\{Y > \lambda+\gamma\} = 0. \qquad \dots \quad (3.7)$$

Now let X be an infinitely divisible random variable with distribution function $F(x)$ for which (i), (ii) and (iii) of Theorem 1 are satisfied. Letting N, M, σ^2, γ now correspond to X it follows that $N(u) = 0$ for $u > 0$, $\sigma^2 = 0$ and

$$\int_{-1}^{0} |x|\,dM(x) < \infty. \qquad \dots \quad (3.8)$$

Let $\epsilon_n > 0$ be such that $\epsilon_n \to 0$ and $-\epsilon_n$ is a continuity point of $M(u)$ for each n. Let X_n be a sequence of independent random variables, with distribution functions $F_n(x)$, associated with $N_n(u) = 0$ for $u > 0$, $\sigma_n^2 = 0$,

$$M_n(u) = \begin{cases} M(u), & u \leqslant -\epsilon_n \\ M(-\epsilon_n), & -\epsilon_n < u < 0 \end{cases}$$

and γ through the formulas (1.1) and (2.1).

Let

$$\lambda_n = \int_{-\infty}^{-\epsilon_n} \frac{|u|}{1+u^2}\, dM(u).$$

From (3.7) with $X_n = Y$ it follows that $P\{X_n > \lambda_n+\gamma\} = 0$. From (3.8) it follows that

$$\lambda_n \to \lambda = \int_{-\infty} \frac{|u|}{1+u^2}\, dM(u) < \infty,$$

and from Theorem 2, of Gnedenko and Kolmogorov [(1954), p. 88] it follows that $F_n(x) \to F(x)$. Thus $P\{X > \lambda+\gamma\} = 0$.

We remark that the above proof furnishes the value of an upper bound for X.

For the proof of Theorem 2 we state without proof the following lemma :

Lemma 5 : *Let X be an infinitely divisible random variable with associated $G_x(u)$, $M_x(u)$, σ_x^2 and $N_x(u)$ given by* (1.1), (2.1) *and* (2.2). *Let $Y=-X$. Then the functions $G_y(u)$, $M_y(u)$, σ_y^2, and $N_y(u)$ given by* 1.1, 2.1 and 2.2 *for the infinitely divisible random variable Y are : $G_y(u)=G_x(+\infty)-G_x(-u)$, $M_y(u)=-N_x(-u)$, $\sigma_y^2=\sigma_x^2$, and* $N_y(u)=-M_x(-u)$.

Theorem 2 follows directly from Lemma 5 and Theorem 1.

4. PROOFS OF THE COROLLARIES AND SOME CONSEQUENCES

In this section we present some particular consequences of Theorems 1 and 2. First, we prove the three corollaries of Section 2.

It follows from (2.2) that if $G(+\infty)>G(-0)$, then either $\sigma^2>0$, or $N(u)\neq 0$ for some $u>0$. In either case, $P\{X>A\}>0$ for any real number A. This demonstrates Corollary 1. To prove Corollary 2, assume that there exists an $\epsilon>0$ such that $G(+\infty)=G(-\epsilon)$. From (2.2) it follows that $M(u)$ is bounded so that conditions (*i*)-(*iii*) of Theorem 1 are satisfied. But the conclusion of Theorem 1, i.e. $P\{X>A\}=0$ for some real number A, contradicts the hypothesis of Corollary 2. Finally, if X is bounded, then $M(u)=0$, $u<0$, $\sigma^2=0$, $N(u)=0\,(u>0)$. In other words $G(u)=0$ for all u, and from (1.1) we have that $\varphi(t)=e^{i\gamma t}$. Thus, $P\{X=\gamma\}=1$, and Corollary 3 is proved.

When applied to certain stochastic processes Theorems 1 and 2 give information about monotonicity of the sample functions. In particular, let $\{x(T),\ 0\leqslant T<\infty\}$ be a separable process with stationary, independent increments for which $x(0)=0$. The Levy-Khintchine representation of $\{x(T),\ 0\leqslant T<\infty)$ is

$$E\{e^{itx(T)}\}=\exp\left\{Ti\gamma t+T\int_{-\infty}^{\infty}\left(e^{itu}-1-\frac{itu}{1+u^2}\right)\frac{1+u^2}{u^2}\,dG(u)\right\}. \quad \dots \quad (3.9)$$

For processes of this type we have the following theorem :

Theorem 4 : *A necessary and sufficient condition that there exists a real number A such that with probability one the functions $y(T)=x(T)+AT$ are non-decreasing is that the $G(u)$ in* (3.9) *satisfies*

(*i*) $G(+0)=G(-\infty)$

(*ii*) $\lim\limits_{\epsilon\to 0}\int_{\epsilon}^{1}\int_{u}^{\infty}\frac{1+x^2}{x^2}\,dG(x)\,du<\infty.$

Moreover, a necessary and sufficient condition that there exists a real number A *such that with probability one the functions* $z(T) = x(T) - AT$ *are non-increasing is that the* $G(u)$ *in* (3.9) *satisfy*

$$(i') \quad G(+\infty) = G(-0)$$

$$(ii') \quad \lim_{\varepsilon \to 0} \int_{-1}^{-\varepsilon} \int_{-\infty}^{u} \frac{1+x^2}{x^2}\, dG(u) du < \infty$$

Proof : We prove only the second half of Theorem 4.

Necessity: Clearly $z(0) = x(0) = 0$ with probability one. If the functions $z(T) = x(T) - AT$ are non-increasing, then $P\{z(1) \leqslant 0\} = 1 = P\{x(1) < A\}$. By Theorem 1 and relations (2.2), (i') and (ii') are easily seen to be satisfied.

Sufficiency: We note from (3.9) that the non-decreasing functions $G_T(u)$, $M_T(u)$, $N_T(u)$, and the constant γ_T associated with the infinitely divisible random variable $x(T)$ through (1.1) and (2.1) are $G_T(u) = TG(u)$, $M_T(u) = TM(u)$, $N_T(u) = TN(u)$ and $\gamma_T = T\gamma$. We note from (2.2) that the conditions of Lemma 4 are satisfied and thus it follows that $P\{x(T) > T(\lambda+\gamma)\} = 0$ where $\lambda = \int_{-\infty}^{0} \frac{|u|}{1+u^2}\, dM(u) < \infty$. Let $A = (\lambda+\gamma)$ so that $z(T) = x(T) - (\lambda+\gamma)T$. Since $x(0) = 0$, the stationarity of the process implies that $P\{z(t) > z(s)\} = 0$ whenever $t > s$. Since the process $z(T)$ is assumed to be separable, it follows that $z(T)$ is monotonic with probability one.

References

Gnedenko, B. V. and Kolmogorov, A. N. (1954) : *Limit Distributions for Sums of Independent Random Variables*, Translated by K. L. Chung, Addison-Wesley.

Pakshirajan, R. P. and Chatterjee, S. D. (1956) : On the unboundedness of infinitely divisible laws. *Sankhyā*, **17**, 349.

Paper received : May, 1959.
Revised : December, 1959.

A LIMIT THEOREM FOR DENSITIES

By R. RANGA RAO

and

V. S. VARADARAJAN

Indian Statistical Institute, Calcutta

SUMMARY. In this paper necessary and sufficient conditions are obtained for the almost everywhere convergence of the density of the normed sum of a sequence of independent and identically distributed random variables, to the density of the standard normal distribution. Extensions to (i) multidimensions and (ii) the case of convergence to stable law are also given. An application to the convergence of conditional distributions is also pointed out.

1. Introduction

Let $\xi_1, \xi_2, \ldots \xi_n, \ldots,$ be a sequence of independent random variables with common distribution function $F(x)$. Let $F_n(x)$ be the distribution function of the normed sums

$$\eta_n = \frac{\xi_1+\xi_2+\ldots+\xi_n}{\sqrt{n}} \qquad \ldots \quad (1.1)$$

and let $p_n(x)$ denote the density of the absolutely continuous component of F_n. It is well-known that $F_n \Longrightarrow \Phi$ ($\Longrightarrow$ denotes weak convergence) if and only if (i) $E\xi_1 = 0$ and (ii) $E\xi_1 = \sigma^2 < \infty$. (See Kolmogorov and Gnedenko (1954), p. 181). Here $\Phi(x)$ denotes the standardized normal distribution. Kolmogorov and Gnedenko in their book (1954, p. 223) have given an example which shows that the density $p_n(x)$ need not 'converge' to $\phi(x)$, the density of $\Phi(x)$. However, an examination of the example reveals that convergence fails to take place only in a neighbourhood of the origin. We are therefore naturally led to ask whether it is true that $p_n(x) \to \phi(x)$ almost everywhere. The object of this note is to answer this question. Theorem 1 shows that almost everywhere convergence can indeed be always asserted.

In this connection, it may be pointed out that complete solutions of the problem are already available for certain other definitions of convergence. Thus, for instance, Gnedenko (1954) has shown that necessary and sufficient conditions that

$$\operatorname*{ess\ sup}_{-\infty < x < +\infty} |p_n(x)-\phi(x)| \to 0 \qquad \ldots \quad (1.2)$$

are (i) $F_n(x) \Longrightarrow \Phi(x)$, and (ii) for some n_0, $F_{n_0}(x)$ satisfies the Lipschitz condition of order 1. Again, Prohorov (1952), (see also Smith, (1953)) has shown that necessary and sufficient conditions that

$$\int_{-\infty}^{\infty} |p_n(x)-\phi(x)|\,dx \to 0 \qquad \ldots \quad (1.3)$$

are (i) $F_n(x) \Longrightarrow \Phi(x)$, and (ii) F_n is non-singular for some n. Now, from (1.3) it at once follows that $p_n(x)$ converges to $\phi(x)$ in measure and this makes it plausible that a.e. convergence takes place.

2. THEOREMS

Theorem 1 : *With the above notation the necessary and sufficient conditions that*

$$p_n(x) \to \phi(x) \qquad \dots \quad (2.1)$$

almost everywhere are (i) $F_n(x) \Longrightarrow \Phi(x)$, *and* (ii) F_n *is non-singular for some* n.

Proof: Necessity. The necessity of (i) follows from Scheffés theorem (1947), if we notice that in view of (2.1), the contribution of the singular component of F_n tends to zero as $n \to \infty$. The necessity of (ii) is obvious.

Sufficiency: Let n_0 be the least integer n for which F_n is non-singular. Such an n_0 exists in view of condition (ii) of the theorem. We can then write $F^{*n_0} = F_A + F_S$ where $F_A(+\infty) = \gamma\,(> 0)$ is the absolutely continuous component of F^{*n_0}. We can therefore choose k so large that if $E = \{x : p(x) < k\}$, $p(x)$ being the density of F_A, then $\int\limits_E p(x)dx = \alpha$ where $0 < \alpha < \gamma$. We can then define

$$G_1(x) = \frac{1}{\alpha} \int\limits_{-\infty}^{x} p(x)\chi_E dx$$

where χ_E is the characteristic function of the set E, and write

$$F^{*n_0} = \alpha G_1 + \beta G_2 \qquad \dots \quad (2.2)$$

where $\beta = 1-\alpha$ and G_1, G_2 are distribution functions. By our choice G_1 is absolutely continuous with bounded density and G_2 is not completely singular. Let $g_1(t)$ and $g_2(t)$ be the c.f.'s (characteristic functions) of the two distribution functions $G_1(x)$ and $G_2(x)$ respectively. Then since G_1 has bounded density it follows that $g_1 \in L_2(-\infty, \infty)$. (See for instance Bochner and Chandrasekharan, (1949), p. 20). Further, since both G_1 and G_2 are non-singular, for any $\delta > 0$ we can find a $\rho = \rho(\delta)$ such that $0 < \rho < 1$ and

$$|g_1(t)| < \rho, \; |g_2(t)| < \rho \qquad \dots \quad (2.3)$$

for all $|t| > \delta$.

Now suppose that $n = n_0k+r$, where k and r are positive integers with $0 \leqslant r < n_0$. Then (* denoting convolution)

$$\begin{aligned} F^{*n} = F^{*n_0k} * F^{*r} &= (\alpha G_1 + \beta G_2)^{*k} * F^{*r} \\ &= \left[\sum_{s=0}^{k} \binom{k}{s} \alpha^s \beta^{k-s} G_1^{*s} * G_2^{*k-s} \right] * F^{*r} \\ &= \left\{ \sum_{s \geqslant 2} \right\} * F^{*r} + \left\{ \sum_{s < 2} \right\} * F^{*r} \\ &= H_n(x) + K_n(x) \quad \text{say.} \end{aligned} \qquad \dots \quad (2.4)$$

It is then clear that $H_n(x)$ is absolutely continuous and let $q_n(x)$ denote the density of $H_n(x\sqrt{n})$. The total mass of K_n is at most $k\,\alpha\,\beta^{k-1}+\beta^k \leqslant C\,\beta^{k/2}$ where C is a constant independent of k. This implies that

$$\int |p_n(x)-q_n(x)|\,dx < C\,\beta^{k/2}. \qquad \ldots \quad (2.5)$$

Now since $0 < \beta < 1$, it follows that $\sum\limits_k \beta^{k/2} < \infty$. By the Borel Cantelli lemma we can then assert that

$$|p_n(x)-q_n(x)| \to 0, \text{ as } n\to\infty \qquad \ldots \quad (2.6)$$

for almost all (Lebesgue) x.

We now proceed to show that $q_n(x)$ tends to $\phi(x)$ uniformly in x. Let $h_n(t)$ denote the c.f. of $H_n(x\sqrt{n})$. Then since $g_1 \epsilon L_2$ and as is easily verified $|h_n(t)| \leqslant |\,g_1(t/\sqrt{n})|^2$ it follows that $h(t) \epsilon L_1$. We then have by the inversion theorem

$$q_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} h_n(t)e^{-itx}dt.$$

Hence

$$|q_n(x)-\phi(x)| \leqslant \frac{1}{2\pi}\int_{-\infty}^{\infty} |\,h_n(t)-e^{-t^2/2}|\,dt$$

$$\leqslant \frac{1}{2\pi}\int_{|t|<A} |h_n(t)-e^{-t^2/2}|\,dt+\frac{1}{2\pi}\int_{A}^{\infty} e^{-t^2/2}dt$$

$$+\frac{1}{2\pi}\int_{A<|t|<\delta\sqrt{n}} |h_n(t)|\,dt+\frac{1}{2\pi}\int_{|t|>\delta\sqrt{n}} |h_n(t)|\,dt$$

$$= I_1+I_2+I_3+I_4 \text{ say.}$$

Since ($f_n(t)$ denoting the c.f. of F_n)

$$|f_n(t)-h_n(t)| \leqslant \beta^k+k\,\alpha\,\beta^{k-1} \qquad \ldots \quad (2.7)$$

and $f_n(t) \to e^{-t^2/2}$ it follows that $h_n(t) \to e^{-t^2/2}$,

that is

$$I_1 = \frac{1}{2\pi}\int_{|t|<A} |h_n(t)-e^{-t^2/2}\,|\,dt = o(1) \qquad \ldots \quad (2.8)$$

for any fixed A. By choosing A sufficiently large I_2 can be made small.

From $F_n(x) \Longrightarrow \Phi(x)$ we can deduce that $F(x)$ has finite second moment. (See Kolmogorov and Gnedenko (1954), p. 181). Hence we can find $\delta > 0$ such that for $|t| \leqslant \delta$ we have

$$|f(t)| \leqslant \exp\left(-\frac{\sigma^2t^2}{4}\right)$$

where σ^2 is the variance. From this and (2.7) it follows that

$$\begin{aligned} I_3 &\leqslant \frac{1}{2\pi} \int\limits_{A<|t|<\delta\sqrt{n}} |f_n(t)|\,dt \\ &\quad + \frac{1}{2\pi}\, 2\delta\sqrt{n}\,(\beta^k + k\alpha\beta^{k-1}) \\ &\leqslant \frac{1}{\pi} \int\limits_A^{\delta\sqrt{n}} e^{-t^2\sigma^2/4}\,dt + o(1) \\ &\leqslant \frac{1}{\pi} \int\limits_A^{\infty} e^{-t^2\sigma^2/4}\,dt + o(1) \qquad \ldots \quad (2.9) \end{aligned}$$

i.e., I_3 can also be made small by choosing A sufficiently large.

Finally from (2.3) it follows that

$$|h_n(t)| \leqslant \rho^{k-2}\, |g_1(t/\sqrt{n})|^2$$

so that $$I_4 \leqslant \rho^{k-2} \int\limits_{|t|>\delta\sqrt{n}} |g_1(t/\sqrt{n})|^2\,dt = \rho^{k-2}\sqrt{n} \int\limits_{|t|>\delta} |g_1(t)|^2\,dt. \qquad \ldots \quad (2.10)$$

i.e., $I_4 = o(1)$ as $n \to \infty$.

Thus combining (2.8), (2.9) and (2.10), we can conclude that

$$|q_n(x) - \phi(x)| \to 0 \qquad \ldots \quad (2.11)$$

uniformly in x. This completes the proof.

By a standard modification of the above argument, Theorem 1 can be generalized to stable distribution functions.

Theorem 2 : *Let* $\xi_1, \xi_2, \ldots \xi_n, \ldots$, *be a sequence of independent and identically distributed random variables, and let*

$$\eta_n = \frac{\xi_1 + \xi_2 + \ldots + \xi_n - A_n}{B_n}.$$

Then the necessary and sufficient conditions that there exist constants A_n *and* B_n (> 0), *such that* $p_n(x) \to \gamma(x)$ *almost everywhere, where* $p_n(x)$ *is the density of the absolutely continuous component of* η_n *and* $\gamma(x)$ *is the density of the stable law* $\Psi(x)$, *are*

(i) *F belongs to the domain of attraction of the stable law* Ψ, *and*

(ii) *F_n is non-singular for some* n.

The proof of the above theorem is entirely analogous to that of Theorem 1 and so is omitted.

3. CONCLUDING REMARKS

With regard to the second condition of the theorem viz., F_n or what is the same thing F^{*n} is non-singular for some n, we may observe that there exist singular distributions F for which F^{*2} is absolutely continuous. (See Salem (1942)). However a characterization of this class of distributions intrinsically in terms of F is not known.

It is worth while to note that Theorem 1 is valid for distributions in finite-dimensional Euclidean spaces. That is, let $X_1, X_2, \ldots$ be a sequence of independent random vectors in k-dimensional Euclidean space E_k with mean zero and variance covariance matrix Σ. Let $p_n(x)$ denote the density function of the distribution F_n of the normed sum $\xi_n = (X_1+X_2+\ldots+X_n)/\sqrt{n}$ and let $\phi(x)$ denote the density of a Normal distribution $\Phi(x)$ with mean vector zero, and variance covariance matrix Σ. Then the necessary and sufficient conditions that $p_n(x) \to \phi(x)$ almost everywhere are (i) $F_n(x) \Longrightarrow \Phi(x)$, and (ii) $F_n(x)$ is non-singular for some n. The proof is the same as in the one-dimensional case.

Let $\Lambda_1(x)$, $\Lambda_2(x)$ be two independent linear functions on E_k. We adopt the same notation as in the above paragraph and assume further, for the sake of simplicity that the common distribution $F(x)$ is absolutely continuous with respect to Lebesgue measure. Then it is clear that the distribution of the two-dimensional random vector $\zeta_n = [\Lambda_1(\xi_n), \Lambda_2(\xi_n)]$ is also absolutely continuous. By central limit theorem the distribution of ζ_n converges weakly to a two-dimensional normal distribution, say $\Phi(\Lambda_1, \Lambda_2)$. Then we can deduce the following theorem from Theorem 1 and its multi-dimensional analogue.

Theorem 3: *Let $p_n(x/y)$ be the density of the conditional distribution of $\Lambda_1(\xi_n)$ given $\Lambda_2(\xi_n)$. Let $\phi(x/y)$ be the corresponding conditional density of the normal distribution $\Phi(\Lambda_1, \Lambda_2)$. Then*

$$p_n(x/y) \to \phi(x/y) \textit{ almost everywhere in } x$$

for almost all (Lebesgue) y. In particular

$$\sup_{A} \, |P_n(A/y) - P(A/y)| \to 0, \textit{ for almost all } y$$

where the supremum is taken over all Borel sets, and $P_n(A/y)$, $P(A/y)$ represent the corresponding conditional distributions.

The above theorem is an immediate consequence of Theorem 1, its multi-dimensional analogue and the fact that if $p_n(x, y)$, $q_n(y)$ denote the densities of ζ_n and $\Lambda_2(\xi_n)$ respectively, then $p_n(x/y) = p_n(x, y)/q_n(y)$. In this connection we might also refer to a paper of Steck (1957).

The almost sure convergence of Theorem 1 can be replaced by uniform convergence, if we impose further assumptions. This, as stated in the introduction has been done by Gnedenko (1954), the assumption required being that F_n be absolutely continuous and $p_n(x)$ be bounded for some n. He also obtains the estimate

$$\underset{-\infty < x < \infty}{\text{ess sup}} \; |p_n(x) - \phi(x)| = 0\left(\frac{1}{\sqrt{n}}\right)$$

if the third moment is finite. In the general case we can still derive estimates for the Lebesgue measure of the set $\{x : |p_n(x) - \phi(x)| > C\}$.

Theorem 4 : *Let* $E|X|^3 < \infty$. *Then there exist constants* C_1, C_2 *and* λ *(possibly depending on the distribution) such that*

$$L\left\{ x : |p_n(x) - \phi(x)| \geqslant \frac{C_1}{\sqrt{n}} \right\} \leqslant C_2 \lambda^n \qquad \dots \quad (3.1)$$

where $0 < \lambda < 1$, *L being Lebesgue measure.*

Proof : With the same notation as in the proof of Theorem 1 we have

$$p_n(x) = q_n(x) + r_n(x)$$

where

$$\int_{-\infty}^{\infty} r_n(x)\, dx \leqslant k\, \alpha\, \beta^{k-1} + \beta^k. \qquad \dots \quad (3.2)$$

Further by standard techniques it is easily shown that

$$\sup_x \; |q_n(x) - \phi(x)| \leqslant \frac{C}{\sqrt{n}} \qquad \dots \quad (3.3)$$

where C is a constant independent of n. In view of this it is clear that

$$E = \left\{ x : |p_n(x) - \phi(x)| > \frac{2C}{\sqrt{n}} \right\} \subset E_1 = \left\{ x : r_n(x) > \frac{C}{\sqrt{n}} \right\}. \qquad \dots \quad (3.4)$$

Thus from (3.2) and (3.4), we deduce

$$L(E) \leqslant L(E_1) \leqslant \frac{\sqrt{n}}{C}\,(k\, \alpha\, \beta^{k-1} + \beta^k). \qquad \dots \quad (3.5)$$

Since k is determined by the relation $n = n_0\, k + r$ $(0 \leqslant r < n_0)$ it is easy to see that (3.1) is an immediate consequence of (3.5). This proves the theorem.

REFERENCES

BOCHNER, S. and CHANDRASEKHARAN, K. (1949): *Fourier Transforms*, Princeton University Press.

GNEDENKO, B. V. (1954): Local limit theorem for densities. *Doklady Akad. Nauk.* SSSR. **95**, 5–7.

KOLMOGOROV, A.N. and GNEDENKO, B. V. (1954): *Limit Distributions for Sums of Independent Random Variables*, Translated by K. L. Chung, Addison-Wesley.

PROHOROV, YU. V. (1952): On a local limit theorem for densities. *Doklady Akad. Nauk.* SSSR, **83**, 797–800.

SALEM, R. (1942): On sets of multiplicity for trigonometric series. *Amer. J. Math.*, **64**, 531-538.

SCHEFFÉ, H. (1947): A useful convergence theorem for probability distributions. *Ann. Math. Stat.*, **18**, 434-438.

SMITH, W. L. (1953): A frequency function form of the central limit theorem. *Proc. Camb. Phil. Soc.*, **49**, 462-472.

STECK, G. P. (1957): *Limit Theorems for Conditional Distributions*, University of California Publications in Statistics, **2**, Number 12.

Paper received : December, 1959.

TESTS FOR INDEPENDENCE AND SYMMETRY IN MULTIVARIATE NORMAL POPULATIONS

By J. ROY

Indian Statistical Institute, Calcutta

SUMMARY. It is shown that the distribution function of each of the statistics proposed by various authors [Mauchly (1940), Votaw (1948), Wilks (1932, 1935, 1946)] for testing independence and symmetry in multivariate Normal population can be expanded in an asymptotic series involving Chi-square probability integrals. For large samples, the first term of the expansion provides a good approximation by itself, the second term being $O(n^{-2})$, where n is the size of the sample. For moderately large samples, the first two terms may be used and the coefficients required for this purpose are tabulated for each one of the various statistics considered.

1. Notation

Given a $p \times p$ matrix $A = ((a_{ij}))$, we shall denote by $\bar{A} = ((\bar{a}_{ij}))$ the $p \times p$ matrix whose elements are

$$\bar{a}_{ii} = a_1, \bar{a}_{ij} = a_2 \text{ if } i \neq j, \qquad \dots \quad (1.1)$$

where

$$a_1 = p^{-1} \sum_{i=1}^{p} a_{ii}, \; a_2 = p^{-1}(p-1)^{-1} \sum_{i \neq j=1}^{p}\sum a_{ij}.$$

Again given a $p \times q$ matrix $B = ((b_{ij}))$, we shall denote by $\bar{\bar{B}} = ((\bar{\bar{b}}_{ij}))$ the $p \times q$ matrix whose elements are

$$\bar{\bar{b}}_{ij} = b, \text{ where } b = (pq)^{-1} \sum_{i=1}^{p} \sum_{j=1}^{q} b_{ij}. \qquad \dots \quad (1.2)$$

Given a $p \times p$ matrix A and a partition $(p_1, p_2, \dots, p_m)$ of p, we shall write A in the partitioned form

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1m} \\ A_{21} & A_{22} & \dots & A_{2m} \\ \dots & \dots & \dots & \dots \\ A_{m1} & A_{m2} & \dots & A_{mm} \end{bmatrix} \qquad \dots \quad (1.3)$$

where $A_{\alpha\beta}$ is a $p_\alpha \times p_\beta$ matrix. Thus writing

$$P_\alpha = p_1 + p_2 + \dots + p_\alpha, \; (\alpha = 1, 2, \dots, m-1); \; P_0 = 0, \; P_m = p \qquad \dots \quad (1.4)$$

we have $A_{\alpha\beta} = ((a_{i_\alpha j_\beta}))$ where the ranges of i_α and j_β are $P_{\alpha-1} < i_\alpha \leqslant P_\alpha$, $P_{\beta-1} < i_\beta \leqslant P_\beta$ for $\alpha, \beta = 1, 2, \ldots, m$. Given a $p \times p$ matrix A and a partition $(p_1, p_2, \ldots, p_m)$ of p we shall denote by A^* the matrix

$$A^* = \begin{bmatrix} \bar{A}_{11} & \bar{\bar{A}}_{12} & \ldots & \bar{\bar{A}}_{1m} \\ \bar{\bar{A}}_{21} & \bar{A}_{22} & \ldots & \bar{\bar{A}}_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ \bar{\bar{A}}_{m1} & \bar{\bar{A}}_{m2} & \ldots & \bar{A}_{mm} \end{bmatrix} \qquad \ldots \quad (1.5)$$

If $p_1 = p_2 = \cdots = p_m$, we shall denote by A^{**} the matrix

$$A^{**} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} & \ldots & \bar{A}_{1m} \\ \bar{A}_{21} & \bar{A}_{22} & \ldots & \bar{A}_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ \bar{A}_{m1} & \bar{A}_{m2} & \ldots & \bar{A}_{mm} \end{bmatrix} \qquad \ldots \quad (1.6)$$

All tests of significance discussed in later sections relate to a p-variate Normal population with means μ_i and dispersion (or, variance-covariance) matrix $((\sigma_{ij}))$, on the basis of a random sample $x_{i\nu}$ drawn from such a population $i, j = 1, 2, \ldots, p$; $\nu = 1, 2, \ldots, n$. The joint probability density function of the $x_{i\nu}$'s is :

$$(2\pi)^{-\frac{1}{2}pn} \, |\sigma_{ij}|^{-\frac{1}{2}n} \exp\Big[-\tfrac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma^{ij} \sum_{\nu=1}^{n} (x_{i\nu}-\mu_i)(x_{j\nu}-\mu_j)\Big] \qquad \ldots \quad (1.7)$$

where $((\sigma^{ij})) = ((\sigma_{ij}))^{-1}$. We shall write

$$\bar{x}_i = n^{-1} \sum_{\nu=1}^{n} x_{i\nu}, \; s_{ij} = \sum_{\nu=1}^{n} (x_{i\nu}-\bar{x}_i)(x_{j\nu}-\bar{x}_j) \text{ and } S = ((s_{ij})). \qquad \ldots \quad (1.8)$$

If the p-variates fall into m groups, with p_α variates in the α-th group, $\alpha = 1, 2, \ldots, m$, $p = p_1+p_2+\ldots+p_m$, without loss of generality, we shall assume that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_m \geqslant 1$, and arrange the variates such that the first p_1-variates belong to the first group, the next p_2-variates belong to the second group and so on and finally the last p_m-variates belong to the m-th group. We shall then write

$$\bar{\bar{x}}_\alpha = p_\alpha^{-1} \sum_{i=P_{\alpha-1}+1}^{P_\alpha} \bar{x}_i. \qquad \ldots \quad (1.9)$$

Further, if the i-th variate belongs to the α-th group and the j-th variate belongs to the β-th group, we shall write

$$t_{ij} = s_{ij}+n(\bar{x}_i-\bar{\bar{x}}_{\alpha})(\bar{x}_j-\bar{\bar{x}}_{\beta}) \qquad \dots \quad (1.10)$$

and $T = ((t_{ij}))$ for the $p\times p$ matrix whose elements are t_{ij}.

The incomplete Beta-function ratio will be denoted by the symbol

$$I_x(P, Q) = \frac{\Gamma(P+Q)}{\Gamma(P)\ \Gamma(Q)} \int\limits_0^x u^{P-1}(1-u)^{Q-1}\,du. \qquad \dots \quad (1.11)$$

2. An asymptotic expansion of a class of distribution functions

Let L be a random variable with Prob $(0 \leqslant L \leqslant 1) = 1$ whose t-th moment about origin is of the form :

$$E(L^t) = \prod_{j=1}^{s} \frac{\left(\frac{n}{2}+b_j\right)_t}{\left(\frac{n}{2}+c_j\right)_t} \qquad \dots \quad (2.1)$$

where n is a positive integer, $c_j > b_j$ $(j = 1, 2, \dots, s)$ are constants not involving n and $(m)_t = \Gamma(m+t)/\Gamma(m)$. It will be seen in later sections that the $\frac{2}{n}$-th power of the various likelihood-ratio statistics for testing hypotheses of independence and symmetry in multivariate Normal populations on the basis of a random sample of size n all have moments of the form (2.1). The probability distribution of specific statistics with moments of this form have been investigated by various authors [Bartlett (1947); Nair (1938); Rao (1948, 1951); Roy (1951); Verma (1951); Wald and Brookner (1941); Wilks and Tukey (1946)]. It is easy to show that asymptotically, for large n, $-n \log_e L$ is distributed as a Chi-square with

$$r = 2\sum_{j=1}^{s}(c_j-b_j) \qquad \dots \quad (2.2)$$

degrees of freedom, which, however, is not a very good approximation when n is only moderately large. The most convenient is the asymptotic series expansion derived by Box (1949) and Roy (1951) which, to terms of order n^{-4} may be expressed as follows :

Let $\quad r_t = \sum\limits_{j=1}^{s}(c_j^t-b_j^t),\ t = 1, 2, 3;\ a = (r_1-r_2)/r_1;\ N = n-a; \qquad \dots \quad (2.3)$

then, writing

$$X = -N\log_e L$$

we have for the probability distribution of X the asymptotic expansion :

$$\text{Prob } (X \geqslant x) = Q_r(x) + \frac{a_2}{N^2}[Q_{r+4}(x) - Q_r(x)] + \frac{a_3}{N^3}[Q_{r+6}(x) - Q_r(x)] + \frac{a_4}{N^4}[Q_{r+8}(x) - Q_r(x)] - \frac{a_2^2}{N^4}[Q_{r+4}(x) - Q_r(x)] + O(N^{-5}) \qquad (2.4)$$

where $r = 2r_1$, is defined by (2.3), $Q_r(x)$ stands for the probability that a Chi-square with r degrees of freedom exceeds x, that is,

$$Q_r(x) = \int_x^\infty \frac{1}{2^{\frac{1}{2}r}\,\Gamma(\frac{1}{2}r)}\, e^{-\frac{1}{2}u}\, u^{\frac{1}{2}r-1}\, du$$

and

$$a_2 = \frac{2}{3}(d_3 - r_1) = \frac{2}{3}(r_3 - r_1) + \frac{1}{2}a(r_1 + r_2) \qquad \ldots \quad (2.5)$$

$$a_3 = 2a_2 - \frac{2}{3}(d_4 - r_1)$$

$$a_4 = 3a_3 - 4a_2 + \frac{1}{2}a_2^2 + \frac{4}{5}(d_5 - r_1)$$

where

$$d_t = \sum_{j=1}^{s}[(c_j + \tfrac{1}{2}a)^t - (b_j + \tfrac{1}{2}a)^t], \; t = 3, 4, 5.$$

The first term of (2.4) by itself provides a good approximation. If the sample size is only moderately large, the first two terms may be used. As it is seldom necessary to use more than the first two terms, we present, for each of the statistics considered in the later sections the values of the constants r, a and a_2 for use in the expansion (2.4) to terms of order N^{-2}.

3. TEST FOR INDEPENDENCE

With a given partition $(p_1, p_2, \ldots, p_m)$ of the p-variates into m groups, to test the hypothesis $\mathcal{H}_i$ that the m groups of variates are mutually independent, Wilks (1935) showed that the likelihood-ratio statistic is the $\frac{1}{2}n$-th power of $L_i = |S| / \prod_{\alpha=1}^{m} |S_{\alpha\alpha}|$ where $S_{\alpha\alpha}$ is the $p_\alpha \times p_\alpha$ diagonal sub-matrix of S defined by (1.8).

Asymptotically, $-n \log_e L_i$ is distributed as a Chi-square with $r = p_1(p_2+p_3+\ldots+p_m)+p_2(p_3+\ldots+p_m)+\ldots+p_{m-1}\,p_m$ degrees of freedom. The exact distribution of L_i in closed form has been obtained by Wald and Brookner (1941) and Wilks (1935) in certain special cases, where the number of groups containing an odd number of variates is atmost one. In the two simplest cases, when $m = 2$ we have for $0 \leqslant x \leqslant 1$,

$$\text{Prob}\quad (L_i \leqslant x) = \begin{cases} I_x\left(\dfrac{n-p+1}{2}, \dfrac{p-1}{2}\right) & \text{if } p_1 = p-1, p_2 = 1 \\ I_{\sqrt{x}}\,(n-p+1, p-2) & \text{if } p_1 = p-2, p_2 = 2 \end{cases} \qquad \dots \quad (3.1)$$

where $I_x(p, q)$ is the incomplete Beta-function defined by (1.11). An infinite series expansion of the distribution of $-n \log_e L_i$ in the general case was obtained by Wald and Brookner (1941).

The t-th moment of L_i computed by Wilks (1935) can be written as

$$E(L_i^t) = \prod_{\alpha=1}^{m-1} \prod_{i_\alpha = P_\alpha + 1}^{P_{\alpha+1}} \frac{\left(\dfrac{n-i_\alpha+1}{2}\right)_t}{\left(\dfrac{n-i_\alpha+1+P_\alpha}{2}\right)_t} \qquad \dots \quad (3.2)$$

where P_α is defined by (1.4). Since the moments are of the form (2.1) the asymptotic expansion (2.4) is available. The constants a, r, a_2 are tabulated for all partitions of p except $(p-1, 1)$ and $(p-2, 2)$ for $3 \leqslant p \leqslant 8$ in columns (4)—(6) of Table 1.

4. Test for spherical symmetry

Mauchly (1940) proposed the hypothesis $\mathcal{H}_s$ of spherical symmetry, namely that the p-variates are mutually independent and have equal variances. He showed that the likelihood-ratio statistic is the $\frac{1}{2}n$-th power of $L_s = |S|/s^p$ where

$$s = \frac{1}{p} \sum_{i=1}^{p} s_{ii}.$$

Asymptotically $-n \log_e L_s$ has the Chi-square distribution with $r = \frac{1}{2}p(p+1)-1$ degrees of freedom. The number of degrees of freedom given by Mauchly (1940) appears to be wrong. The exact probability distribution of L_s when $p = 2$ is worked out by Mauchly (1940) as

$$\text{Prob}\quad (L_s \leqslant x) = x^{\frac{1}{2}(n-2)} \text{ for } p = 2,\ 0 \leqslant x \leqslant 1. \qquad \dots \quad (5.1)$$

The t-th moment of L_s about origin computed by Mauchly (1940) can be expressed as

$$E(L_s^t) = \prod_{i=1}^{p-1} \frac{\left(\dfrac{n-i-1}{2}\right)_t}{\left(\dfrac{n-1}{2} + \dfrac{i}{p-1}\right)_t} \qquad \dots \quad (5.2)$$

which is of the form (2.1). The constants a, r, a_2 in the asymptotic expansion (2.4) for $3 \leqslant p \leqslant 8$ are tabulated in columns (2)—(4) of Table 2.

5. TEST FOR EQUALITY OF MEANS, EQUALITY OF VARIANCES AND EQUALITY OF COVARIANCES

5.1. *The hypothesis* $\mathcal{H}_{vc}$ *:* Wilks (1946) showed that the likelihood-ratio statistic for testing the hypothesis $\mathcal{H}_{vc}$ that the variances are equal and the covariances are equal is the $\frac{1}{2}n$-th power of $L_{vc} = |S|/|\bar{S}|$ where $\bar{S}$ is derived from S by formula (1.1).

Asymptotically $-n \log_e L_{vc}$ follows the Chi-square distribution with $r = \frac{1}{2}p(p+1)-2$ degrees of freedom. The exact probability distributions for the cases $p = 2$, 3 obtained by Wilks (1946) are as follows :

$$\text{Prob}\quad (L_{vc} \leqslant x) = \begin{cases} I_x\left(\dfrac{n-2}{2}, \dfrac{1}{2}\right) \text{ for } p = 2 \\[2ex] I_{\sqrt{x}}\,(n-3,\ 2) \quad \text{for } p = 3 \end{cases} \qquad \ldots \quad (5.1)$$

for $0 \leqslant x \leqslant 1$. Verma (1951) has obtained infinite series expansions for the distribution of L_{vc} for other small values of p.

The t-th moment of L_{vc} derived by Wilks (1946) is given by

$$E(L_{vc}^t) = \prod_{i=1}^{p-1} \frac{\left(\dfrac{n-1+i-p}{2}\right)_t}{\left(\dfrac{n-1}{2}+\dfrac{i-1}{p-1}\right)_t} \qquad \ldots \quad (5.2)$$

which is of the form (2.1). The constants a, r, a_2 in the asymptotic expansion (2.4) are tabulated in columns (5)—(7) of Table 2 for $4 \leqslant p \leqslant 8$.

5.2. *The hypothesis* $\mathcal{H}_{mvc}$ *:* For testing the hypothesis $\mathcal{H}_{mvc}$ that the means are equal, the variances are equal and the covariances are equal, Wilks (1946) showed that the likelihood-ratio criterion is the $\frac{1}{2}n$-th power of $L_{mvc} = |S|/|\bar{V}|$ where $\bar{V}$ is derived from the matrix $V = ((v_{ij}))$ by formula (1.1) where $v_{ij} = s_{ij}+n(\bar{x}_i-\bar{\bar{x}})(\bar{x}_j-\bar{\bar{x}})$ and $\bar{\bar{x}} = \dfrac{1}{p}\sum\limits_{i=1}^{p} \bar{x}_i$.

Asymptotically, $-n \log_e L_{mvc}$ is distributed as Chi-square with $r = \frac{1}{2}p(p+3)-3$ degrees of freedom. The exact probability distributions for $p = 2$, 3 are as follows :

$$\text{Prob}\quad (L_{mvc} \leqslant x) = \begin{cases} I_x\left(\dfrac{n-2}{2}, 1\right) \text{ for } p = 2 \\[2ex] I_{\sqrt{x}}\,(n-3,\ 3) \text{ for } p = 3 \end{cases} \qquad \ldots \quad (5.3)$$

An infinite series expansion of the distribution of L_{mvc} for other small values of p has been obtained by Verma (1951).

The t-th moment of L_{mvc} computed by Wilks (1946) is

$$E(L^t_{mvc}) = \prod_{i=1}^{p-1} \frac{\left(\frac{n+i-p-1}{2}\right)_t}{\left(\frac{n}{2}+\frac{i-1}{p-1}\right)_t} \qquad \dots \quad (5.4)$$

which is of the form (2.1). The asymptotic expansion (2.4) is therefore available and the constants a, r, a_2 involved are tabulated in columns (8)–(10) of Table 2 for $4 \leqslant p \leqslant 8$.

6. TEST FOR COMPOUND SYMMETRY

In whatever follows, unless anything to the contrary is explicitly stated, we shall consider a partition $(p_1, p_2, \dots, p_m)$ of the p-variates into m groups. If there are q groups of variates each containing just one variate, we shall write $h = m-q$ $(h \geqslant 1)$ so that $p_\alpha \geqslant 2$ for $\alpha = 1, 2, \dots h$ and $p_\alpha = 1$ for $\alpha = h+1, h+2, \dots, m$. For brevity, we shall denote such a partition by $(p_1, \dots p_h, 1^q)$. We are now in a position to state various hypotheses of compound symmetry proposed by Votaw (1948) who obtained the respective likelihood-ratio statistics and their moments.

6.1. *The hypothesis* $\mathscr{H}_{1(vc)}$: This hypothesis states that within each group of variates, the variances are equal and the covariances are equal, and between groups, the covariances are equal for each pair of groups. The likelihood-ratio criterion is the $\frac{1}{2}n$-th power of $L_{1(vc)} = |S|/|S^*|$ where S^* is derived from S by the formula (1.5) corresponding to the partition $(p_1, p_2, \dots, p_h, 1^q)$ of p.

Asymptotically $-n \log_e L_{1(vc)}$ has the Chi-square distribution with $r = \frac{1}{2}[p(p+1)-q(q+1)-h(h+3)]-qh$ degrees of freedom. The exact probability distribution of $L_{1(vc)}$ has been worked out in Votaw (1948) for the partitions $(2, 1^{p-2})$ and $(3, 1^{p-3})$ of p. Thus :

$$\text{Prob}\quad (L_{1(vc)} \leqslant x) = \begin{cases} I_x\left(\frac{n-p}{2}, \frac{p-1}{2}\right) \text{ for the partition } (2, 1^{p-2}) \\ I_{\sqrt{x}}\,(n-p,\ p-1) \text{ for the partition } (3,\ 1^{p-3}). \end{cases} \qquad \dots \quad (6.1)$$

Some other special cases are derived by Votaw, Kimball and Rafferty (1950).

With P_α defined by (1.4), the t-th moment of $L_{1(vc)}$ can be expressed as follows as shown by Votaw (1948)

$$E(L^t_{1(vc)}) = \prod_{\alpha=1}^{h} \prod_{i_\alpha=1}^{p_\alpha-1} \frac{\frac{1}{2}(n-m-i_\alpha-P_{\alpha-1}+\alpha-1)_t}{\left(\frac{n-1}{2}+\frac{i_\alpha-1}{p_\alpha-1}\right)_t} \qquad \ldots \quad (6.2)$$

which is of the form (2.1). The values of the constants a, r, a_2 in the asymptotic expansion (2.4) are tabulated in columns (7)—(9) of Table 1 for all partitions of p except $(2, 1^{p-2})$ and $(3, 1^{p-3})$ for $4 \leqslant p \leqslant 8$.

6.2. *The hypothesis* $\mathscr{H}_{1(mvc)}$: This states that within each group of variates the means are equal, the variances are equal and the covariances are equal and between groups, the covariances are equal for each pair of groups. The likelihood-ratio statistic is the $\frac{1}{2}$ n-th power of $L_{1(mvc)} = |S|/|T^*|$ where the matrix T is defined by formula (1.10) for the partition $(p_1, p_2, \ldots, p_h, 1^q)$ of the p-variates and T^* is derived from T by formula (1.5).

Asymptotically $-n \log_e L_{1(mvc)}$ is distributed as a Chi-square with $r = \frac{1}{2}[p(p+3)-q(q+3)-h(h+5)]-hq$ degrees of freedom. The exact probability distribution of $L_{1(mvc)}$ for the partitions $(2, 1^{p-2})$ and $(3, 1^{p-3})$ obtained by Votaw (1948) are as follows :

$$\text{Prob } (L_{1(mvc)} \leqslant x) = \begin{cases} I_x\left(\frac{n-p}{2}, \frac{p}{2}\right) \text{ for the partition } (2, 1^{p-2}) \\ I_{\sqrt{x}}\,(n-p,\ p) \text{ for the partition } (3, 1^{p-3}) \end{cases} \qquad \ldots \quad (6.3)$$

and certain other cases are examined by Votaw, Kimball and Rafferty (1950). The t-th moment of $L_{1(mvc)}$ computed by Votaw (1948) is

$$E(L^t_{1(mvc)}) = \prod_{\alpha=1}^{h} \prod_{i_\alpha=1}^{p_\alpha-1} \frac{\frac{1}{2}(n-m-i_\alpha-P_{\alpha-1}+\alpha-1)_t}{\left(\frac{n}{2}+\frac{i_\alpha-1}{p_\alpha-1}\right)_t} \qquad \ldots \quad (6.4)$$

where P_α is defined in (1.4). This is of the form (2.1) and the constants a, r, a_2 involved in the asymptotic expansion (2.4) are given in columns (10)—(12) of Table 1 for all partitions of p except $(2, 1^{p-2})$ and $(3, 1^{p-3})$ for $4 \leqslant p \leqslant 8$.

In the special case, $p_1 = p_2 = \ldots = p_m = u$, $u \geqslant 2$, that is when the p-variates fall into m groups, each of u-variates, for a fixed arrangement of the u-variates within each group, two other hypotheses of compound symmetry have been considered by Votaw (1948). These are discussed in the following sections.

6.3. *The hypothesis* $\mathscr{H}_{1(vc)}$: This hypothesis states that within each group, the variances are equal and the covariances are equal and between groups, the diagonal covariances are equal and the off-diagonal covariances are equal. The likelihood-ratio criterion is the $\frac{1}{2}n$-th power of $\bar{L}_{1(vc)} = |S|/|S^{**}|$ where S^{**} is derived from S by formula (1.6) corresponding to the partition $(u, u, \ldots, u)$ of $p = mu$.

Asymptotically $-n \log_e \bar{L}_{1(vc)}$ follows the Chi-square distribution with $r = \frac{1}{2}p(p+1) - m(m+1)$ degrees of freedom. In the case where $m = 2$ and $u = 2$, the exact distribution of $\bar{L}_{1(vc)}$ derived in Votaw (1948) is as follows:

$$\text{Prob } (\bar{L}_{1(vc)} \leqslant x) = I_{\sqrt{x}}(n-4,\, 2) \text{ when } u = 2,\, m = 2. \qquad \ldots \quad (6.5)$$

The t-th moment of $\bar{L}_{1(vc)}$ was shown to be

$$E(\bar{L}^t_{1(vc)}) = \prod_{\alpha=1}^{m} \prod_{i_\alpha=1}^{u-1} \frac{\left(\dfrac{n-m-i_\alpha-(u-1)(\alpha-1)}{2}\right)_t}{\left(\dfrac{n-1}{2}+\dfrac{1-\alpha}{2(u-1)}+\dfrac{i_\alpha-1}{u-1}\right)_t} \qquad \ldots \quad (6.6)$$

which is of the form (2.1) and the values of the constants a, r, a_2 in the asymptotic expansion (2.4) are presented in columns (4)—(6) of Table 3 for $m = 2$, $u = 3, 4$; $m = 3$, $u = 2$, and $m = 4$, $u = 2$.

6.4. *The hypothesis* $\mathscr{H}_{1(mvc)}$: This hypothesis specifies that within each group, the means are equal, the variances are equal and the covariances are equal and between groups, the diagonal covariances are equal and the off-diagonal covariances are equal. The likelihood-ratio criterion was shown (Votaw, 1948) to be the $\frac{1}{2}n$-th power of $\bar{L}_{1(mvc)} = |S|/|T^{**}|$ where the matrix T is defined by (1.10) corresponding to the partition $(u, u, \ldots, u)$ of $p = mu$ and T^{**} is derived from T by means of formula (1.6).

Asymptotically $-n \log_e \bar{L}_{1(mvc)}$ is distributed as a Chi-square with $r = \frac{1}{2}p(p+3) - m(m+2)$ degrees of freedom. The exact distribution in the case $m = 2$, $u = 2$ obtained by Votaw (1948) is as follows:

$$\text{Prob } (\bar{L}_{1(mvc)} \leqslant (x)) = I_{\sqrt{x}}(n-4,\, 3) \text{ when } u = 2,\, m = 2. \qquad \ldots \quad (6.7)$$

The t-th moment of $\bar{L}_{1(mvc)}$ as derived by Votaw (1948) is :

$$E(\bar{L}^t_{1(mvc)}) = \prod_{\alpha=1}^{m} \prod_{i_\alpha=1}^{u-1} \frac{\left(\dfrac{n-m-i_\alpha-(u-1)(\alpha-1)}{2}\right)_t}{\left(\dfrac{n}{2}+\dfrac{1-\alpha}{2(u-1)}+\dfrac{i_\alpha-1}{u-1}\right)_t} \qquad \ldots \quad (6.8)$$

This is of the form (2.1) and the values of the constants a, r, a_2 in the asymptotic expansion (2.4) are presented in columns (7)—(9) of Table 3 for $m = 2$, $u = 3, 4$, $m = 3$, $u = 2$ and $m = 4$, $u = 2$.

TABLE 1. VALUES OF CONSTANTS a, r, a_2 FOR L_i, $L_{1(mvc)}$ AND $L_{1(vc)}$ CRITERIA

p	m	partition	independence L_i			compound symmetry					
						$L_{1(mvc)}$			$L_{1(vc)}$		
			a	r	a_2	a	r	a_2	a	r	a_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
3	2	2,1									
	3	1,1,1	1.83333	3	-.02083	*	*	*	*	*	*
4	2	3,1									
	2	2,2				2.78571	7	1.41964	3.30000	5	.38750
	3	2,1,1	2.30000	5	.38750						
	4	1,1,1,1	2.16667	6	.45833	*	*	*	*	*	*
5	2	4,1				2.70635	14	6.86486	3.21717	11	3.48141
	2	3,2				2.88462	13	5.45673	3.40000	10	2.60000
	3	3,1,1	2.78571	7	1.41964						
	3	2,2,1	2.75000	8	1.37500	3.27778	9	3.32639	3.78571	7	4.86696
	4	2,1,1,1	2.61111	9	1.65972						
	5	1,1,1,1,1	2.50000	10	1.87500	*	*	*	*	*	*
6	2	5,1				2.96429	21	14.36830	3.48529	17	8.24908
	2	4,2				3.09444	20	12.62207	3.61806	16	7.08314
	2	3,3	3.50000	9	2.43750	3.10000	20	12.45000	3.62500	16	6.93750
	3	4,1,1	3.27778	9	3.32639	3.19935	17	11.99778	3.70635	14	6.86486
	3	3,2,1	3.22727	11	3.35796	3.37500	16	9.93750	3.88462	13	5.45673
	3	2,2,2	3.16667	12	3.41667	3.56667	15	7.79583	4.08333	12	3.97917
	4	3,1,1,1	3.08333	12	3.97917	*	*	*	*	*	*
	4	2,2,1,1	3.03846	13	3.99519	3.77273	11	6.35795	4.27778	9	3.32639
	5	2,1,1,1,1	2.92857	14	4.48214						
	6	1,1,1,1,1,1	2.83333	15	4.89583	*	*	*	*	*	*
7	2	6,1				3.24483	29	26.16543	3.77500	24	16.29625
	2	5,2				3.34821	28	24.02623	3.88043	23	14.79280
	2	4,3	4.00000	12	5.00000	3.35317	28	23.79354	3.88647	23	14.59256
	3	5,1,1	3.77273	11	6.35796	3.45000	25	23.10938	3.96429	21	14.36830
	3	4,2,1	3.71429	14	6.71429	3.57870	24	20.65728	4.09444	20	12.62207
	3	3,3,1	3.70000	15	6.66250	3.58333	24	20.45833	4.10000	20	12.45000
	3	3,2,2	3.62500	16	6.93750	3.71739	23	18.04076	4.23684	19	10.73355
	4	4,1,1,1	3.56667	15	7.79583	3.69444	20	19.25540	4.19935	17	11.99778
	4	3,2,1,1	3.50000	17	7.93750	3.86842	19	16.41776	4.37500	16	9.93750
	4	2,2,2,1	3.44444	18	8.11111	4.05556	18	13.48611	4.56667	15	7.79583
	5	3,1,1,1,1	3.38889	18	8.81944						
	5	2,2,1,1,1	3.34211	19	8.94408	4.26923	13	10.76442	4.77273	11	6.35795
	6	2,1,1,1,1,1	3.25000	20	9.68750						
	7	1,1,1,1,1,1,1	3.16667	21	10.35417	*	*	*	*	*	*
8	2	7,1				3.53801	38	43.58350	4.07639	32	28.78665
	2	6,2				3.62432	37	40.99453	4.16452	31	26.89024
	2	5,3	4.50000	15	9.06250	3.62838	37	40.72255	4.16935	31	26.65272
	2	4,4	4.50000	16	9.00000	3.62913	37	40.67215	4.17025	31	26.60870
	3	5,1,1,	4.26923	13	10.76442	3.72353	34	39.65029	4.24483	29	26.16543
	3	5,2,1	4.20588	17	11.81985	3.82576	33	36.74953	4.34821	28	24.02623
	3	4,3,1	4.18421	19	11.83882	3.82997	33	36.48370	4.35317	28	23.79354
	3	4,2,2	4.10000	20	12.45000	3.93403	32	33.68740	4.45885	27	21.74551
	3	3,3,2	4.07143	21	12.47321	3.93750	32	33.46875	4.46296	27	21.55324
	4	5,1,1,1	4.05556	18	13.48611	3.93966	29	34.97360	4.45000	25	23.10938
	4	4,2,1,1	3.97619	21	13.99702	4.06746	28	31.69037	4.57870	24	20.65728
	4	3,3,1,1	3.95455	22	13.98864	4.07143	28	31.46429	4.58333	24	20.45833
	4	3,2,2,1	3.89130	23	14.43207	4.20370	27	28.21991	4.71739	23	17.87409
	4	2,2,2,2	3.83333	24	14.83333	4.34615	26	24.72115	4.86364	22	15.39773
	5	4,1,1,1,1	3.86364	22	15.39773	4.19082	23	29.01285	4.69444	20	19.25540
	5	3,2,1,1,1	3.79167	24	15.73958	4.36364	22	25.27273	4.86842	19	16.41776
	5	2,2,2,1,1	3.74000	25	16.07750	4.54762	21	21.42560	5.05556	18	13.48611
	6	3,1,1,1,1,1	3.70000	25	16.93750						
	6	2,2,1,1,1,1	3.65385	26	17.22115	4.76667	15	16.79583	5.26923	13	10.76442
	7	2,1,1,1,1,1,1	3.57407	27	18.27546						
	8	1,1,1,1,1.1,1,1	3.50000	28	19.25000	*	*	*	*	*	*

TABLE 2. VALUES OF a, r, a_2 FOR L_s, L_{vc} AND L_{mvc} CRITERIA

p	spherical symmetry L_s			equality of means, of variances and of covariances					
				L_{vc}			L_{mvc}		
	a	r	a_2	a	r	a_2	a	r	a_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
3	1.27778	5	0.45910						
4	1.58333	9	1.85938	2.73611	8	1.47184	2.21718	11	3.48140
5	1.90000	14	4.86500	3.01923	13	4.24880	2.48530	17	8.24908
6	2.22222	20	10.30864	3.32105	19	9.41040	2.77500	24	16.29624
7	2.54762	27	19.19005	3.63248	26	17.95537	3.07638	32	28.78664
8	2.87500	35	32.67578	3.94958	34	31.04982	3.38502	41	47.05200

TABLE 3. VALUES OF a, r, a_2 FOR $\bar{L}_{1(vc)}$ AND $\bar{L}_{1(mvc)}$ CRITERIA

p	m	partition	compound symmetry					
			$\bar{L}_{1(vc)}$			$\bar{L}_{1(mvc)}$		
			a	r	a_2	a	r	a_2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
6	2	(3,3)	3.75000	15	39.34896	3.22368	19	44.34567
6	3	(2,2,2)	4.50000	9	2.43750	4.00000	12	5.00000
8	2	(4,4)	4.25926	30	89.72119	3.71605	36	37.54596
8	4	(2,2,2,2)	5.50000	16	9.00000	5.00000	20	15.00000

An asterisk in the above tables indicate that the hypothesis is not defined, a blank that the exact distribution is available in closed form.

REFERENCES

BARTLETT, M. S. (1947): Multivariate analysis. *J. Roy. Stat. Soc.* (Sup.), **9**, 76.

BOX, G. E. P. (1949): A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317.

MAUCHLY, J. W. (1940): Significance test for sphericity of normal distribution *Ann. Math. Stat.*, **11**, 204.

NAIR, U. S. (1938): Application of the moment function in the study of distribution laws in statistics. *Biometrika*, **30**, 274.

RAO, C. R. (1948): Tests of significance in multivariate analysis. *Biometrika*, **35**, 58.

——— (1951): An asymptotic expansion of the distribution of Wilks' Λ criterion. *Bull. Int. Stat. Inst.*, **33**, 177.

ROY, J. (1951) : The distribution of certain likelihood-ratio criteria useful in multivariate analysis. *Bull. Int. Stat. Inst.*, **33**, part 2, 219.

——— (1954) : On some tests of significance in samples from bi-polar normal distributions. *Sankhya*, **14**, 203.

VERMA, K. B. (1951) : The distribution of Wilks' L_{mvc} and L_{vc} criteria. *Bull. Inst. Stat. Int.*, **33**, part 2, 181.

VOTAW, D. F. JR. (1948) : Testing compound symmetry in a normal multivariate distribution. *Ann. Math. Stat.*, **19**, 447.

VOTAW, D. F. JR., KIMBALL, A. R. and RAFFERTY, J. A. (1950) : Compound symmetry tests in multivariate analysis of medical experiments. *Biometrics*, **6**, 259.

WALD, A. and BROOKNER, J. R. (1941) : On the distribution of Wilks' statistic for testing the independence of several groups of variates. *Ann. Math. Stat.*, **12**, 137.

WILKS, S. S. (1932) : Certain generalisations in the analysis of variance. *Biometrika*, **24**, 471.

——— (1935) : On the dependence of k sets of normally distributed statistical variables. *Econometrica*, **3**, 309.

——— (1946) : Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution. *Ann. Math. Stat.*, **17**, 257.

WILKS, S. S. and TUKEY, J. W. (1946) : Approximation of the distribution of the product of Beta-variables by a single Beta-variable. *Ann. Math. Stat.*, **17**, 318.

Paper received : February, 1960.

ON THE F-TEST IN THE INTRABLOCK ANALYSIS OF A BALANCED INCOMPLETE BLOCK DESIGN

By SUJIT KUMAR MITRA

Indian Statistical Institute, Calcutta

SUMMARY. In this paper the distribution of the criterion (treatment sum of squares)/(treatment+error sum of squares) is considered over possible random assignments of treatments to the experimental plots. Computing the first two moments of this distribution for the situation when treatment effects are absent and comparing them with the corresponding moments of a Beta distribution it is concluded that, when block errors are homogeneous and the total number of experimental plots is moderately large, the usual variance ratio test is expected to give a sufficiently accurate evaluation of the level of significance of the randomisation test.

1. INTRODUCTION

In a balanced incomplete block experiment the v treatments are assigned to the experimental plots arranged in b blocks of k plots each in such a way that (a) each treatment occurs exactly once in r blocks, (b) each pair of treatments occurs together equally often (λ times) within some block. The treatment arrangement in the corresponding plan follows exactly the same rule. However, we may conveniently call the blocks of the plan as 'groups' to distinguish them from the experimental blocks. The randomisation procedure in a BIBD can now be described as follows.

R_1 : Firstly the b groups in the plan are assigned to the b experimental blocks completely at random.

R_2 : The treatments in a group are then randomly assigned to plots within a block separately and independently for each block.

We denote the yield with treatment $m(=1, 2, \ldots, v)$ on plot $j(=1, 2, \ldots, k)$ of block $i(=1, 2, \ldots, b)$ by X_{ijm}.

The treatments are said to be equal in effects if the observed yield of a treatment on a plot is purely a plot effect and is independent of the treatment, that is, if,

$$H_0: \quad X_{ij1} = X_{ij2} = \ldots = X_{ijv} = X_{ij}, \qquad \text{for each} \quad i = 1, 2, \ldots v; \quad j = 1, 2, \ldots k,$$

Let the treatment m occur in groups $i_1, i_2, \ldots, i_r$ of the plan ($i_1 < i_2 < \ldots < i_r$) and let the observed yields to treatment m, in whichever blocks these groups happen to be assigned, be denoted by $x_{1m}, x_{2m}, \ldots, x_{rm}$ respectively. Denote by B_i the total yield of block i and by $\sum\limits_{(m)} B_i$ the total yield of all the r blocks in which treatment m appears. Write

$$T_m = \sum_{h=1}^{r} x_{hm}$$

$$Q_m = T_m - \sum_{(m)} B_i/k$$

$$G = \sum_{i=1}^{b} B_i.$$

The ANOVA table for the intrablock analysis of a BIBD is obtained (Yates, 1936) as follows.

TABLE 1. ANALYSIS OF VARIANCE FOR A BALANCED INCOMPLETE BLOCK DESIGN

sources of variation	d.f.	s.s.	m.s.
blocks (ignoring treatments)	$b-1$	$S_{b+t} = \Sigma B_i^2/k - C.F.$	
treatments (eliminating blocks)	$v-1$	$S_{t\cdot b} = \dfrac{k}{\lambda v} \Sigma Q_m^2$	$M_{t\cdot b}$
error	$bk-b-v+1$	S_e (by subtraction)	M_e
total	$bk-1$	$\Sigma\Sigma x_{hm}^2 - C.F.$	

$C.F. = G^2/bk$

The null hypothesis H_0 is tested by the randomisation test using the criterion $M_{t.b}/M_e$. This is evaluated for the layout actually used and also for all other layouts that could have been obtained under R_1 and R_2. (Note that this is possible, under H_0, since then each plot would record the same yield, namely what it has presently recorded, irrespective of the treatment applied). If the value of this criterion for the layout actually used is exceeded in α percent of the possible arrangements we say that it is significant at the α percent level. This is a laborious task and one naturally looks for approximations.

The object of the present paper is to investigate the extent to which the usual variance ratio distribution $F(\overline{v-1}/\overline{bk-b-v+1})$ can be used to approximate the distribution of the criterion over possible arrangements in the same way as Welch (1937) and Pitman (1938) investigated this for the Randomised Block Design and Welch (1937) did for the Latin Square. This is done by evaluating the moments of a related criterion $S_{t.b}/(S_{t.b}+S_e)$ and comparing them with the corresponding moments of a Beta distribution $B\left(\frac{v-1}{2}, \frac{bk-b-v+1}{2}\right)$.

2. THE MEAN AND VARIANCE OF $S_{t.b}/(S_{t.b}+S_e)$ UNDER H_0

Put

$$x'_{hm} = x_{hm} - \text{mean yield of the corresponding block}$$

$$e_{ij} = X_{ij} - \bar{X}_{i.} \quad \text{where} \quad \bar{X}_{i.} = \frac{1}{k} \sum_j X_{ij},$$

$$S_i = \sum_j e_{ij}^2, \qquad S = \sum_i S_i,$$

$$U = \sum_{ij} e_{ij}^4. \qquad \dots (2.1)$$

We note that

$$Q_m = \sum_h x'_{hm} \qquad \dots (2.2)$$

and that, under H_0

$$S_{t.b} + S_e = S \qquad \dots (2.3)$$

which is thus non-stochastic.

We also note that for any fixed assignment under R_1, the x'_{hm}s coming from different groups are distributed independently of one another and that if a particular group is thus assigned to block i the possible values for the vector of x'_{hm} for the k treatments which occur in that group are precisely the $k!$ permutations of $(e_{i1}, e_{i2}, \dots, e_{ik})$. Any x' here would thus have conditionally a mean value of zero and variance $\frac{S_i}{k}$.

Hence

$$E\,(\sum_m Q_m^2) = \sum_m E(Q_m^2) = \sum_m E[E(Q_m^2 \mid R_1) = vE[E(Q_1^2 \mid R_1)] = \frac{rv}{bk} \cdot S = S.$$

Therefore

$$E(S_{t.b}/(S_{t.b}+S_e)) = \frac{k}{\lambda v} S/S = \frac{k}{\lambda v} = \frac{k(v-1)}{v\lambda(v-1)} = \frac{v-1}{b(k-1)}, \qquad \dots (2.4)$$

since $\lambda(v-1) = r(k-1)$.

We have similarly

$$E(Q_1^4) = E[E(Q_1^4 \mid R_1)]$$

$$= r\left(\frac{U}{bk}\right)+\frac{3r(r-1)(S^2-\Sigma S_i^2)}{k^2b(b-1)}.$$

$$E(Q_1^2Q_2^2) = E[E(Q_1^2Q_2^2 \mid R_1)]$$

$$= \frac{(r^2-\lambda)(S^2-\Sigma S_i^2)}{k^2b(b-1)}+\frac{\lambda(\Sigma S_i^2-U)}{bk(k-1)}+\frac{2\lambda(\lambda-1)(S^2-\Sigma S_i^2)}{k^2(k-1)^2b(b-1)}.$$

Hence

$$E(\Sigma\, Q_m^2)^2 = vE(Q_1^4)+v(v-1)E(Q_1^2\, Q_2^2)$$

$$= \frac{rv}{bk}U+\frac{\lambda v(v-1)}{bk(k-1)}(\Sigma\, S_i^2-U)+$$

$$+\left[3rv(r-1)+(r^2-\lambda)v(v-1)+\frac{2\lambda(\lambda-1)v(v-1)}{(k-1)^2}\right]\frac{(S^2-\Sigma\, S_i^2)}{k^2b(b-1)}$$

$$= S^2+rv\left[3(r-1)+r(v-1)-(k-1)+\frac{2(\lambda-1)}{k-1}-k(b-1)\right]\frac{(S^2-\Sigma\, S_i^2)}{k^2b(b-1)}$$

$$= S^2+\frac{2rv}{k-1}\left[(r-1)(k-1)+(\lambda-1)\right]\frac{(S^2-\Sigma\, S_i^2)}{k^2b(b-1)}$$

$$= S^2+\frac{2rv(\lambda v-k)}{k^2b(b-1)(k-1)}\left[\left(1-\frac{1}{b}\right)S^2-\Sigma(S_i-\bar{S})^2\right]$$

$$= S^2+\frac{2rv(\lambda v-k)}{k^2b^2(k-1)}\,S^2\left[1-\frac{b\Sigma(S_i-\bar{S})^2}{(b-1)S^2}\right].$$

Therefore $V(\Sigma\, Q_m^2) = S^2.\, \dfrac{2rv(\lambda v-k)}{k^2b^2(k-1)}\left[1-\dfrac{\delta}{b-1}\right]$

where $$\delta = \frac{V(S_i)}{\bar{S}^2} = \left[\frac{\text{coef. of variation in the } S_i\text{'s}}{100}\right]^2 . \qquad \ldots \quad (2.5)$$

We have thus $$V\left[\frac{S_{t.b}}{S_e+S_{t.b}}\right] = \frac{2rv(\lambda v-k)}{\lambda^2v^2b^2(k-1)}\left[1-\frac{\delta}{b-1}\right]. \qquad \ldots \quad (2.6)$$

3. COMPARISON WITH THE MOMENTS OF THE BETA DISTRIBUTION

The mean and variance of $B\left(\frac{v-1}{2}, \frac{bk-b-v+1}{2}\right)$ are given by

$$\text{mean} = \frac{v-1}{b(k-1)} \qquad \ldots \quad (3.1)$$

$$\text{variance} = \frac{2(v-1)(bk-b-v+1)}{[b(k-1)]^2[b(k-1)+2]}$$

$$= \frac{2(v-1)(bk-b-v+1)}{[b(k-1)]^3} \text{ if } b(k-1) \gg 2$$

$$= \frac{2rv(\lambda v-k)}{\lambda^2 v^2 b^2(k-1)}.$$

Thus while the mean value of $S_{t.b}/(S_{t.b}+S_e)$ agrees perfectly with the mean of $B\left(\frac{v-1}{2}, \frac{bk-b-v+1}{2}\right)$ the variances would also be practically equal if $b(k-1) \gg 2$ and the block errors are homogeneous, that is, $\delta = 0$. Here, possibly the F-test works reasonably well as an approximate randomisation test.

Heterogeneity in block errors reduces the variance of $S_{t.b}/(S_{t.b}+S_e)$. Hence if δ is large a reference to the suggested Beta distribution for testing H_0, or equivalently the usual F-test for $M_{t.b}/M_e$, would generally underestimate the significance of departure from H_0, and would thus be somewhat conservative.

The situation is thus remarkably similar to that of a randomised block design and in some sense possibly better. Note that the ratio of the variance of $S_{t.b}/(S_{t.b}+S_e)$ to that of the corresponding Beta distribution is approximately equal to $1-\frac{\delta}{b-1}$ in a BIBD and to $1-\frac{\delta}{r-1}$ in a randomised block design. In a BIBD, $b > r$, and in many cases b is very large compared to r.

The types of departures from the null hypothesis H_0 that can possibly be detected by the analysis of variance F-test have been examined by Rao (1959) through a comparison of the non-null expected values of the treatment and error mean squares.

References

Pitman, E. J. G. (1938): Significance tests which may be applied to samples from any population, III. The analysis of variance test. *Biometrika*, **29**, 332-335.

Rao, C. R. (1959): Expected values of mean squares in the analysis of incomplete block experiments and some comments based on them. *Sankhyā*, **21**, 327-336.

Welch, B. L. (1937): On the z-test in the randomised blocks and Latin Squares. *Biometrika*, **29**, 21–52.

Yates, F. (1936): A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci.*, **26**, 424-455.

Paper received : November, 1959.

AN ADMISSIBLE ESTIMATE FOR ANY SAMPLING DESIGN

By V. P. GODAMBE*

Indian Statistical Institute, Calcutta

SUMMARY. In the class of all unbiassed estimates, admissibility of a well-known estimate is established.

1. Introduction

The author (1955) defined the general sampling design, and the corresponding class of linear unbiassed estimates, for finite populations, as follows: Let

$$1, \ldots, \lambda, \ldots, N \qquad \ldots \quad (1.1)$$

denote the different individuals in the population, the corresponding variate values being

$$X_1, \ldots, X_\lambda, \ldots, X_N. \qquad \ldots \quad (1.2)$$

The problem is to estimate

$$T = \sum_{\lambda=1}^{N} X_\lambda \qquad \ldots \quad (1.3)$$

by observing X values of a few individuals λ, from (1.1). Any sequence

$$s = \lambda_1, \ldots, \lambda_{n_s} \qquad \ldots \quad (1.4)$$

of (not necessarily distinct) individuals from (1.1) is called a 'sample' and be denoted by s. Further, let

$$S = \{s\} \qquad \ldots \quad (1.5)$$

be an arbitrary 'finite' set of sequences s in (1.4). For every $s \varepsilon S$ we define a non-negative number $P_s \geqslant 0$, such that

$$\sum_{s \varepsilon S} P_s = 1. \qquad \ldots \quad (1.6)$$

Now if we put

$$P = \{P_s\} \quad s \varepsilon S \qquad \ldots \quad (1.7)$$

a 'sampling design' d is defined as

$$d = (S, P). \qquad \ldots \quad (1.8)$$

It is easy to see that if D denotes the class of all sampling designs d in (1.8), then all the known 'sample survey designs' must belong to D. In fact, for every $d \varepsilon D$ it is possible to construct a sampling mechanism of drawing individuals from (1.1), 'one after another' (Hanumantha Rao, 1960).

---

* Now with Science College, Nagpur, India.

For any given sample $s \varepsilon S$, we define a linear estimate e_s as

$$e_s = \sum_{\lambda \varepsilon s} \beta_{s\lambda} \, . \, X_\lambda, \qquad \dots \quad (1.9)$$

where the summation is taken over all the 'distinct' individuals λ in s. It is again clear that all the known linear estimates must be particular cases of e_s in (1.9). Now, for unbiassedness of e_s,

$$E(e_s) = T \qquad \dots \quad (1.10)$$

for all T in (1.3). A necessary and sufficient condition for e_s in (1.9) to be unbiassed, in a sampling design d, is

$$\sum_{s \ni \lambda} \beta_{s\lambda} \, . \, P_s = 1, \quad (\lambda = 1, \dots, N) \qquad \dots \quad (1.11)$$

where $s \ni \lambda$, stands 'for all s which include λ'. Further if e_s is unbiassed, its variance is given by,

$$V(e_s) = \sum_{\lambda=1}^{N} X_\lambda^2 \sum_{s \ni \lambda} \beta_{s\lambda}^2 \; P_s + \sum_{\lambda \neq \lambda'} X_\lambda X_{\lambda'} \sum_{s \ni \lambda, \lambda'} \beta_{s\lambda} \beta_{s\lambda'} P_s - T^2. \qquad \dots \quad (1.12)$$

2. An admissible estimate

The probability of any particular individual λ being included in the sample is given by

$$\sum_{s \ni \lambda} P_s = P(\lambda) \qquad \dots \quad (2.1)$$

for $\lambda = 1, \dots, N$. Following Hájek, we call

$$\bar{e}_s = \sum_{\lambda \varepsilon s} X_\lambda / P(\lambda) \qquad \dots \quad (2.2)$$

a simple linear estimate (Hájek, 1959). It is easy to check from (1.11) that $\bar{e}_s$ is unbiassed. We call an unbiassed estimate e_s 'admissible' if for any other unbiassed estimate e_s'

$$V(e_s) < V(e_s') \qquad \dots \quad (2.3)$$

for some value of $X = (X_1, \dots, X_\lambda, \dots X_N)$ in (1.2). It is proved below that in this sense the simple linear estimate $\bar{e}_s$ in (2.2) is admissible.

Let

$$e_s' = \sum_{\lambda \varepsilon s} \beta_{s\lambda}' X_\lambda \qquad \dots \quad (2.4)$$

be an unbiassed estimate. If e_s' is different from $\bar{e}_s$ in (2.2) it follows that

$$\beta_{s\lambda}' \neq 1/P(\lambda) \qquad \dots \quad (2.5)$$

for some λ and s, $\lambda \varepsilon s$. To be specific, let us suppose for $\lambda_0 \varepsilon s_o$

$$\beta_{s_0\lambda_0}' \neq 1/P(\lambda_0). \qquad \dots \quad (2.6)$$

Now if $X_{\lambda_0} = 1$ and $X_\lambda = 0$ for $\lambda \neq \lambda_0$ in (1.12), we have

$$V(\bar{e}_s) = \frac{1}{P(\lambda_0)} - 1, \qquad \ldots \quad (2.7)$$

and

$$V(e'_s) = \sum_{s \ni \lambda_0} \beta'^2_{s\lambda_0} P_s - 1, \qquad \ldots \quad (2.8)$$

from (1.12). From (2.7), (2.8) and (1.11) we have,

$$V(e'_s) - V(\bar{e}_s) = \sum_{s \ni \lambda_0} \left(\beta'_{s\lambda_0} - \frac{1}{P(\lambda_0)}\right)^2. \qquad \ldots \quad (2.9)$$

The r.h.s. of (2.9) is > 0 because of (2.6). Hence $\bar{e}_s$ in (2.2) is admissible.*

3. Illustrations

First consider the sampling mechanism of making a fixed number, say n, of draws, with replacement and with equal probabilities. Then for the resulting sampling design we have from (2.1),

$$P(\lambda) = 1 - \left(1 - \frac{1}{N}\right)^n. \qquad \ldots \quad (3.1)$$

The simple linear estimate (2.2) in this case is given by

$$\bar{e}_s = \sum_{\lambda \epsilon s} X_\lambda / 1 - \left(1 - \frac{1}{N}\right)^n. \qquad \ldots \quad (3.2)$$

Letting $v(s)$ = number of distinct individuals in s, we get another unbiassed estimate,

$$e'_s = N \sum_{\lambda \epsilon s} X_\lambda / v(s). \qquad \ldots \quad (3.3)$$

It follows from the admissibility of $\bar{e}_s$ in (3.2) that

$$V(\bar{e}_s) < V(e'_s) \qquad \ldots \quad (3.4)$$

for some value of $X = (X_1, \ldots, X_\lambda, \ldots, X_N)$ in (1.2).

It has been proved (Basu, 1958) that the estimate e'_s in (3.3) has uniformly smaller variance than the conventional arithmetic mean. However, (3.4) proves that e'_s cannot be a best unbiassed estimate. In fact, it has been demonstrated (Godambe, 1955) that in the whole class of linear unbiassed estimates of the population total, a uniformly minimum variance estimate does not exist.

Next, consider sampling with replacement and with equal probabilities, until distinct individuals are sampled, where v is given in advance. In this case

$$P(\lambda) = v/N, \quad \lambda = 1, \ldots, N, \qquad \ldots \quad (3.5)$$

since, for the general sampling design, if $v(s)$ denote the number of distinct individuals in the sample s,

$$E(v(s)) = \sum_1^N P(\lambda). \qquad \ldots \quad (3.6)$$

---

* By an independent argument, earlier the author (Godambe, 1955) specified the class of prior distributions, with respect to which, e_s is the Bayes solution.

It follows from (2.2) and (3.5) that an admissible estimate in the present case is

$$\bar{e}_s = N \sum_{\lambda \epsilon s} X_\lambda / v. \quad \ldots \quad (3.7)$$

4. An admissible estimate which minimises maximum variance

From (1.12) and (2.2) we have,

$$V(\bar{e}_s) = \sum_{\lambda=1}^{N} X_\lambda^2 . \frac{1}{P(\lambda)} + \sum_{\lambda \neq \lambda'} X_\lambda X_{\lambda'} . \frac{P(\lambda, \lambda')}{P(\lambda)P(\lambda')} - T^2. \quad \ldots \quad (4.1)$$

If we assume that $X = (X_1, \ldots, X_\lambda, \ldots, X_N)$ in (1.2) is such that

$$X_\lambda \geqslant 0, \quad \lambda = 1, \ldots, N \quad \ldots \quad (4.2)$$

we have (since $P(\lambda, \lambda')/P(\lambda) \leqslant 1$)

$$V(\bar{e}_s) \leqslant T \sum_{1}^{N} \frac{X_\lambda}{P(\lambda)} - T^2 \leqslant T^2 . \left(\frac{1}{P(\lambda_M)} - 1 \right), \quad \ldots \quad (4.3)$$

where $P(\lambda_M) =$ minimum of $\{P(1), \ldots, P(\lambda), \ldots, P(N)\}$. ... (4.4)

Thus, for any sampling design, (4.3) gives an upper bound for $V(\bar{e}_s)$, for all X_s's satisfying (4.2). Moreover, for any given T, this upper bound is actually attained when $X_{\lambda_M} = T$ and $X_\lambda = 0$, $\lambda \neq \lambda_M$, in which case

$$V(\bar{e}_s) = T^2 \left(\frac{1}{P(\lambda_M)} - 1 \right). \quad \ldots \quad (4.5)$$

Now, the maximum variance in (4.5) is minimised for a sampling design for which $P(\lambda_M)$ in (4.4) is maximum. If we restrict ourselves to the sampling designs for which the expected number, $E(v(s))$, of distinct individuals in a sample s, is fixed, the design for which $P(\lambda_M)$ is maximum, or (4.5) is minimum, is immediately suggested from the fact that

$$E(v(s)) = \sum_{1}^{N} P(\lambda) \quad \ldots \quad (4.6)$$

whatever the sampling design may be. Thus, for the sampling design obtained by drawing a fixed number, $E(v(s)) = n$ (say) of individuals with equal probabilities, and without replacement, $P(\lambda_M)$ in (4.4) is maximised, and then the maximum variance in (4.5) is minimised by

$$\bar{e}_s = N \sum_{\lambda \epsilon s} X_\lambda / n. \quad \ldots \quad (4.7)$$

References

Aggrawal, O. M. P. (1959): Bayes and minimax procedures in sampling from finite and infinite populations. *Ann. Math. Stat.*, **30**, No. 1.

Basu, D. (1958): On sampling with and without replacement. *Sankhyā*, **20**, 287-94.

Godambe, V. P. (1955): A unified theory of sampling from finite populations. *J. Roy. Stat. Soc.*, Series B, **17**, No. 2, 269-278.

Hájek, J. (unpublished): An optimum strategy and other problems in probability sampling.

Rao, Hanumanth (1960) : An existense theorem in sampling. (Read before Indian Science Congress, 1960).

Paper received : September, 1959.

NON-NULL DISTRIBUTION OF THE LIKELIHOOD-RATIO IN ANALYSIS OF DISPERSION

By J. ROY

Indian Statistical Institute, Calcutta

SUMMARY. In this paper the non-null distribution of Wilks' likelihood-ratio criterion for analysis of dispersion (multivariate analysis of variance) when the expectation-matrix is of unit rank is worked out in a form suitable for numerical evaluation of the power function when the deviation parameter is small and degrees of freedom for error moderately large. An illustrative table of the power function of the analysis of dispersion test at five percent level of significance is presented for $p = 1, 2, 3, 4$ variates when the degrees of freedom for the hypothesis are $m = 2, 3$ and the degrees of freedom for error are $n = 200$.

1. Introduction

It is well known that problems of analysis of dispersion (or, multivariate analysis of variance) can be reduced to the following form: The joint probability density function of the elements of the random matrices X of form $p \times n (n \geqslant p)$ and Y of form $p \times m$ is

$$(2\pi)^{\frac{1}{2}p(n+m)} \; |\Sigma|^{-\frac{1}{2}(n+m)} \; \exp \left[-\tfrac{1}{2} \text{ tr } \Sigma^{-1} \{XX' + (Y-M)(Y-M)'\} \right] \quad \ldots (1.1)$$

where the matrix Σ of form $p \times p$ is positive-definite and unknown and the problem is to test the hypothesis H_0 that the matrix M of form $p \times m$ is a null-matrix, that is,

$$H_0 : \quad M = 0 \quad \ldots (1.2)$$

The likelihood-ratio statistic for testing this hypothesis is due to Wilks (1932) and can be put as

$$L = \frac{|XX'|}{|XX' + YY'|} \, . \quad \ldots (1.3)$$

The sampling distribution of this statistic in the null case when H_0 is true has been investigated extensively; see Rao (1953) or Anderson (1957). In the null case, we shall say that L follows Wilks' distribution with degrees of freedom n, m and p and denote its probability density function by $W(L, n, m, p)$. In this paper, we shall be concerned with the non-null distribution of L when $M \neq 0$. The following properties of the null-distribution of L would be useful in this connection:

$$E(L^t | H_0) = \prod_{i=0}^{p-1} \frac{\left(\frac{n-i}{2}\right)_t}{\left(\frac{n+m-1-i}{2}\right)_t} \quad \ldots (1.4)$$

where

$$(a)_t = \Gamma(a+t)/\Gamma(a).$$

Under H_0, *the statistic* L *is distributed independently of the elements of* $(XX' + YY')$. ... (1.5)

The non-null distribution of L when $m = 1$ was derived by Bose and Roy (1938) and Hsu (1938) and the probability density function in this case can be written as :

$$\sum_{j=0}^{\infty} p_j(\tfrac{1}{2}\delta^2)\, B\left(L, \frac{n-p+1}{2}, \frac{p}{2}+j\right) \qquad \dots \quad (1.6)$$

where $p_j(\theta)$ is the Poisson probability function :

$$p_j(\theta) = e^{-\theta}\, \theta^j/j! \qquad \dots \quad (1.7)$$

and $B(L, r, s)$ is the Beta density function

$$B(L, r, s) = \frac{1}{B(r, s)} L^{r-1}(1-L)^{s-1} \qquad \dots \quad (1.8)$$

and

$$\delta^2 = M'\Sigma^{-1}M. \qquad \dots \quad (1.9)$$

Anderson (1946) has shown that, in general, when the rank of the matrix M is q, $q \leqslant \min(p, m)$, the distribution of L can involve at most the q parameters $\delta_1, \delta_2, \dots, \delta_q$ defined as the positive square-roots of the non-zero roots of the determinantal equation

$$|MM'-\delta^2\Sigma| = 0. \qquad \dots \quad (1.10)$$

He has also derived the moments of L when $t = 1, 2$.

2. A LEMMA

To obtain the non-null distribution of L we make the linear transformation

$$[X^* \vdots Y^*] = C_1B[X \vdots Y]C_2 \qquad \dots \quad (2.1)$$

where the matrix B of the form $p\times p$ satisfies

$$B\Sigma^{-1}B' = I \qquad \dots \quad (2.2)$$

and orthogonal matrices C_1 of the form $p\times p$ and C_2 of the form $m\times m$ are chosen to make

$$C_1\, BM\, C_2 = \left[\begin{array}{c:c} & O_2 \\ O_1 & \cdots\cdots \\ & \Delta \end{array}\right] \qquad \dots \quad (2.3)$$

where O_1 and O_2 are null-matrices of the forms $p\times(m-q)$ and $(p-q)\times q$ respectively and Δ is a $q\times q$ diagonal matrix with diagonal elements $\delta_1, \delta_2, \dots, \delta_q$ defined by (1.10). The existence of matrices C_1 and C_2 is proved by Deemer and Olkin (1951). Writing Y^* in the partitioned form

$$Y^* = [Y_1^* \vdots Y_2^*] \qquad \dots \quad (2.4)$$

where Y_1^* is of the form $p\times(m-q)$ and Y_2^* is of the form $p\times q$, it follows that

$$L = \frac{|X^*X^{*\prime}|}{|X^*X^{*\prime}+Y_1^*Y_1^{*\prime}+Y_2^*Y_2^{*\prime}|} \qquad \dots \quad (2.5)$$

and that the joint probability density function of X^*, Y_1^*, Y_2^* is

$$(2\pi)^{-\frac{1}{2}p(n+m)} \exp\left[-\tfrac{1}{2}\operatorname{tr}\left\{X^*X^{*\prime}+Y_1^*Y_1^{*\prime}+\left(Y_2^*-\begin{bmatrix}O_2\\ \Delta\end{bmatrix}\right)\left(Y_2^*-\begin{bmatrix}O_2\\ \Delta\end{bmatrix}\right)'\right\}\right]. \qquad \dots \quad (2.6)$$

Now let

$$L_1 = \frac{|X^*X^{*\prime}|}{|X^*X^{*\prime}+Y_1^*Y_1^{*\prime}|}, \quad L_2 = \frac{|Z_1Z_1'|}{|Z_1Z_1'+Z_2Z_2'|} \qquad \dots \quad (2.7)$$

where

$$Z_1 = [X^* : Y_1^*], \qquad Z_2 = Y_2^* \qquad \dots \quad (2.8)$$

so that

$$L = L_1 L_2. \qquad \dots \quad (2.9)$$

Obviously L_1 follows Wilks' distribution with n, $m-q$ and p degrees of freedom and from (1.5) it is easy to see that the distribution of L_1 is independent of L_2. We thus have the

Lemma 2.1: *The probability distribution of the statistic L defined by* (1.3) *is the same as the product of two independent statistics L_1 and L_2 where L_1 follows Wilks' distribution with degrees of freedom n, $m-q$ and p and L_2 is defined as the determinantal ratio* (2.7) *where the joint probability density function of the elements of Z_1 of form $p\times(n+m-q)$ and Z_2 of form $p\times q$ is*

$$(2\pi)^{-\frac{1}{2}p(n+m)} \exp\left[-\tfrac{1}{2}\operatorname{tr}\{Z_1Z_1'+(Z_2-M_2)(Z_2-M_2)'\}\right] \qquad \dots \quad (2.10)$$

where $M_2 = \begin{bmatrix}O_2\\ \Delta\end{bmatrix}$.

The problem of deriving the distribution of L_2 is simpler than that of L because the value of m is now reduced to only q.

3. THE CASE $q = 1$

In this case, the probability density function of L_2 is

$$\sum_{j=0}^{\infty} p_j(\tfrac{1}{2}\delta^2)B\left(L_2, \frac{n+m-p}{2}, \frac{p}{2}+j\right) \qquad \dots \quad (3.1)$$

where $p_j(\theta)$ is defined by (1.7), and δ is the single parameter involved. Also, L_1 follows independently Wilks' distribution with degrees of freedom n, $m-1$ and p. Consequently, the probability that the product $L = L_1L_2$ is less than a preassigned constant x, $0 < x < 1$, may be evaluated as

$$\text{Prob }(L < x) = \sum_{j=0}^{\infty} p_j(\tfrac{1}{2}\delta^2)P_j(x) \qquad \dots \quad (3.2)$$

where

$$P_j(x) = \int\limits_{L_1L_2<x} W(L_1, n, m-1, p)B\left(L_2, \frac{n+m-p}{2}, \frac{p}{2}+j\right)dL_1dL_2. \qquad \dots \quad (3.3)$$

To evaluate $P_j(x)$ we note that the problem is the same as finding the cumulative distribution function of the statistic

$$L_{0j} = L_1 \,.\, L_{2j} \qquad \dots \quad (3.4)$$

where L_1 and L_{2j} are distributed independently, L_1 having the probability density function $W(L_1, n, m-1, p)$ and L_{2j} having the probability density function $B\left(L_2, \frac{n+m-p}{2}, \frac{p}{2}+j\right)$. Consequently the t-th moment of L_{0j} is given by

$$E(L_{0j}^t) = E(L_{2j}^t)E(L_1^t) = \frac{\left(\frac{n}{2}+\frac{m-p}{2}\right)_t}{\left(\frac{n}{2}+j\right)_t} \cdot \prod_{i=0}^{p-1} \frac{\left(\frac{n}{2}-\frac{i}{2}\right)_t}{\left(\frac{n}{2}+\frac{m-1-i}{2}\right)_t} \qquad \dots \quad (3.5)$$

where $(a)_t$ is as defined in (1.4).

It is known, however, (Roy, 1951), that for any statistic V distributed in the interval (0, 1) with t-th moment about origin given by

$$E(V^t) = \prod_{j=1}^{s} \frac{\left(\frac{n}{2}+b_j\right)_t}{\left(\frac{n}{2}+c_j\right)_t} \qquad \dots \quad (3.6)$$

where $c_j > b_j$, $j = 1, 2, \dots, s$ are constants not involving n,

$$\text{Prob}\,(V < x) = Q_r(x^*) + \text{terms } O(n^{-2}) \qquad \dots \quad (3.7)$$

where

$$x^* = -(n-\lambda)\log_e x, \quad r = 2r_1, \quad \lambda = (r_1-r_2)/r_1 \qquad \dots \quad (3.8)$$

$$r_t = \sum_{j=1}^{s} (c_j^t - b_j^t) \quad t = 1, 2 \qquad \dots \quad (3.9)$$

and $Q_r(x)$ is the probability for a Chi-square with r degrees of freedom to exceed x, that is,

$$Q_r\,(x) = \int_x^\infty \frac{1}{2^{\frac{1}{2}r}\Gamma(\frac{1}{2}r)}\, e^{-\frac{1}{2}u} u^{\frac{1}{2}r-1} du. \qquad \dots \quad (3.10)$$

Thus

$$P_j(x) = Q_{pm+2j}(X_j) + \text{ terms } O(n^{-2}) \qquad \dots \quad (3.11)$$

where

$$X_j = -(n+m-\lambda_j)\log_e x \qquad \dots \quad (3.12)$$

$$\lambda_j = \frac{pm(p+m+1)+2j-4j^2}{2pm+2j}. \qquad \dots \quad (3.13)$$

Formulae (3.2) and (3.12) give a very convenient method for computation when n is large and δ small. Since

$$|\operatorname{Prob}(L < x) - \sum_{j=0}^{k} p_j(\tfrac{1}{2}\delta^2)P_j(x)| < 1 - \sum_{j=0}^{k} p_j(\tfrac{1}{2}\delta^2) \qquad \ldots (3.14)$$

and when δ is small, even for comparatively small values of k the difference $1 - \sum_{j=0}^{k} p_j(\frac{1}{2}\delta^2)$ is rather small, we need consider only a very few terms in the expansion (3.2).

If n is so large that terms $O(n^{-2})$ may be neglected, $P_j(x)$ can be computed from Tables of the incomplete Gamma function or from Hartley and Pearson's (1950) Tables of the Chi-square integral. The p_j's can be read off from Molina's (1943) Tables of Poisson probabilities.

4. Power function of analysis of dispersion test

Formula (3.2) can be used to evaluate the power function of the analysis of dispersion test when the alternative hypothesis is of rank one. To illustrate, consider the case where there are $p = 4$ variates, the degrees of freedom for error and hypotheses are $n = 200$ and $m = 2$ respectively, the level of significance is fixed at five percent ($\alpha = 0.05$) and the alternative hypothesis of unit rank specifies that $\delta^2 = 2$. Using Rao's (1951) approximation (see Rao, 1953) to the null distribution of L, namely that $-\left(n + \frac{m-p-1}{2}\right) \times \log_e L$ follows the Chi-square distribution with pm degrees of freedom, the lower five percent point of the distribution of L is found to be

$$x = 0.92486.$$

The power of the test is thus given by Prob $(L < x)$. To compute this, we prepare first the following table of values of $pm+2j$, λ_j and X_j for $j = 0, 1, 2, 3, 4, 5$:

j	$pm+2j$	λ_j	X_j
0	8	3.5	15.5073
1	10	3.0	15.5464
2	12	2.2	15.6089
3	14	1.181818	15.6884
4	16	0	15.7807
5	18	−1.307692	15.8829

Writing $p_j = p_j(\frac{1}{2}\delta^2)$ and $Q_j = Q_{pm+2j}(X_j)$ we read off the values of p_j and Q_j from Hartley and Pearson's (1950) Tables :

j	p_j	Q_j
0	0.36788	0.05000
1	0.36788	0.11338
2	0.18394	0.20981
3	0.06131	0.33278
4	0.01533	0.46836
5	0.00307	0.60070

The power of the test is thus given by

$$\text{Prob } (L < x) = \Sigma\, p_j Q_j = 0.12812 \simeq 0.13$$

to two places of decimals.

The following table gives the power of analysis of dispersion tests at the five percent level of significance for $p = 1, 2, 3, 4,\quad m = 2, 3,\quad n = 200,\quad \delta^2 = 0(1)8$.

POWER OF ANALYSIS OF DISPERSION TEST WHEN THE ALTERNATIVE HYPOTHESIS IS OF UNIT RANK

$n = 200$ $\alpha = .05$

	probability of rejection of the hypothesis							
	$m = 2$				$m = 3$			
δ^2	number of variates (p)				number of variates (p)			
	1	2	3	4	1	2	3	4
0	.05	.05	.05	.05	.05	.05	.05	.05
1	.13	.10	.09	.09	.11	.09	.08	.08
2	.22	.16	.14	.13	.19	.14	.12	.11
3	.31	.23	.20	.17	.26	.20	.17	.15
4	.40	.31	.26	.23	.35	.26	.22	.19
5	.49	.38	.33	.28	.43	.33	.27	.23
6	.57	.45	.39	.34	.50	.39	.32	.28
7	.64	.52	.45	.40	.57	.45	.38	.33
8	.70	.59	.52	.46	.64	.52	.43	.38

References

Anderson, T. W. (1957): *Introduction to Multivariate Analysis*, John Wiley and Sons.

Bose, R. C. and Roy S. N. (1938): The exact distribution of the Studentised D^2-statistic. *Sankhyā*, **4**, 19.

Deemer, W. L. and Olkin, I. (1951): The Jacobians of certain matrix transformations useful in multivariate analysis. *Biometrika*, **38**, 345.

Hartley and Pearson (1950): Tables of χ^2-integral and the cumulative Poisson distribution. *Biometrika*, **37**, 313.

Hsu, P. L. (1938) : Notes on Hotelling's generalised *T*. *Ann. Math. Stat.*, **9**, 231.

Rao, C. R. (1953): *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons.

Roy, J. (1951): The distribution of certain likelihood-ratio criteria useful in multivariate analysis. *Proc. Int. Stat. Inst.*, **33**(2), 219.

Wilks, S. S. (1932): Certain generalisations in the analysis of variance. *Biometrika*, **24**, 471.

Paper received : December, 1959.

NONPARAMETRIC LINEAR ESTIMATION OF COMMON MEDIAN OF SYMMETRICAL POPULATIONS FROM SYMMETRICALLY CENSORED OBSERVATIONS

JOHN E. WALSH

System Development Corporation, Santa Monica, California, USA

SUMMARY. Estimation of the mean of a specified continuous parametric population on the basis of a linear function of some or all of the order statistics of a sample from that population has received the attention of several authors. The case of a doubly censored sample has been considered by A. E. Sarhan, who gave some approximate rules for efficient selection of the coefficients in the linear estimate based on the available order statistics. This paper considers a nonparametric application of these rules to the situation of a symmetrically censored sample that is known to have come from a symmetrical population (not necessarily continuous). The detailed shape of this population is not known but is believed to be, at least roughly, of a specified parametric form. On the basis of Sarhan's rules, a nonparametric estimate is obtained for the population median that has a reasonably high efficiency if the observations are a sample and the population sampled is at least approximately of the stated type. This estimate furnishes nonparametric protection in that it can be valid when the observations are not a sample and some or all of the populations yielding the observations are not even roughly of the specified shape. Specifically, the procedure furnishes an estimate for the common median when the observations are independent and from any symmetrical populations having the same central value. For the special case of no censoring, the estimation statistics can be used to obtain permutation tests which are reasonably efficient when the observations are a symmetrically censored sample and the shape of the symmetrical population is at least roughly known. The same type of nonparametric protection is furnished as for the estimation case. Some large sample approximations are given for use in applying these tests. Also, a method is stated for obtaining confidence intervals on the basis of these tests.

INTRODUCTION

The estimates and tests presented are intended for application in situations where the observations (before censoring) are believed to be a random sample and the population distribution is believed to be symmetrical with a shape that is at least roughly known. The statistical procedures presented are reasonably efficient when these assumptions hold and furnish protection against some of the possible assumption violations. The rules presented for determining the statistics used are applicable for a moderately comprehensive spectrum of distribution shapes; consequently, these procedures can be applied to a wide variety of situations involving samples from symmetrical populations.

The underlying data consist of n observations whose ordered values are denoted by $x(1) \leqslant x(2) \leqslant \ldots \leqslant x(n)$. The available set of observations is symmetrically censored so that only the values of $x(r)$, $x(r+1)$, ..., $x(n-r)$, $x(n+1-r)$ are known, where $0 \leqslant r \leqslant (n+1)/2$. The linear estimates considered are of the form

$$\hat{\varphi} = \hat{\varphi}[a(r), \ldots, a(n+1-r)] = \sum_{i=r}^{n+1-r} a(i)x(i), \qquad \ldots \quad (1)$$

where
$$a(i) = a(n+1-i), \quad \sum_{i=r}^{n+1-r} a(i) = 1.$$

This paper furnishes some rules for selecting suitable values for the $a(i)$ when the type of situation considered arises.

The method given for determining these coefficients is based on two papers by A. E. Sarhan (1955a and 1955b), who developed some approximate selection rules for the case of doubly censored samples from continuous parametric populations. The version of these rules presented here is in terms of probability density functions, but their application is not restricted to the continuous case. These rules can be applied to the discrete case on the basis of the shapes of histograms for discrete frequency functions.

Following Sarhan's general approach, the $a(i)$ are divided into three sets such that all of the coefficients in a set have the same value. An estimate is determined (for given r and n) by specifying the size of each set and the coefficient value for each set.

Although $\hat{\varphi}$ was developed for the purpose of estimating the median φ of a somewhat specialized population, its validity as an estimate is rather general. If the only requirement is that the observations are statistically independent and from n possibly different symmetrical populations with common central value φ, the statistic $\hat{\varphi}$ can be shown to have a distribution that is symmetrical about φ. This $\hat{\varphi}$ is a median estimate of φ under rather general conditions, and, if $E(\hat{\varphi})$ exists, is also an unbiased estimate of φ.

One situation involving symmetrical censoring is that where no observations are removed (i.e., $r = 1$). For this case, permutation (randomization) tests can be obtained for φ on the basis of $\hat{\varphi}$; also confidence intervals can be determined on the basis of these tests. Since $\hat{\varphi}$ has a reasonably high estimate efficiency (in the sense that the variance for $\hat{\varphi}$ is not much larger than the minimum variance for a linear estimate of φ), these tests and confidence intervals can be expected to be reasonably efficient. For large sample sizes, the test statistic has an approximately normal distribution, so that the efficiency level is directly established as being the same as that for estimation. For moderate and small sample sizes, the efficiency level is difficult to determine; however, intuitive considerations indicate that it should be reasonably high.

Unless n is quite small, the direct application of one of these permutation tests can require much effort. When the sample size is at least moderately large, however, a normal approximation that is not difficult to apply can frequently be used.

The use of $\hat{\varphi}$ for permutation tests has the same general validity as the use of $\hat{\varphi}$ as an estimate. Namely, the only requirement for the tests to be valid is that the observations are independent and from n possibly different symmetrical populations with common central value φ.

The next section contains a statement of the rules for determining the $a(i)$. The following section contains the results for permutation tests and confidence intervals. The final section contains verification of the statement that $\hat{\varphi}$ is symmetrically distributed about φ under rather general conditions.

Selection of coefficients

As already mentioned, the coefficients are divided into three sets such that all the $a(i)$ in a set have the same value. Let

$x(z) = x$ (largest integer $\leqslant z+1/2$), $a(z) = a$ (largest integer $\leqslant z+1/2$). Then the three sets are determined by (1) and the additional relations

$$a(r+1) = \ldots = a[(n+1+r)/A], \quad a[1+(n+1+r)/A] = \ldots = a[(n+1)/2].$$

where

$$2(n+1+r)/(n-1) \leqslant A \leqslant (n+1+r)/(r+1).$$

Here it is to be noted that one set consists of $a(r)$ and $a(n+1-r)$. Thus the three sets and the corresponding $a(i)$ values are determined by specifying A and any two of $a(r)$, $a(r+1)$, $a[(n+1)/2]$.

The approximate rules for use in selecting values for $a(r), \ldots, a(n+1-r)$ are

(a) For a U-shaped population, $a(r+1) = 0$, $a(r)$ is large and positive. A increases as the shape approaches that of a rectangular distribution.

(b) For a rectangular population, $a(r+1) = a[(n+1)/2] = 0$.

(c) For a triangular or parabolic shaped unimodal distribution, all $a(i) \geqslant 0$, $a(r) = a(r+1)$, $a(r+1) \geqslant a[(n+1)/2]$, and A decreases as population shape tends toward that for normality (develops tails).

(d) For a normal population, $a(r) = a(r+1) = a[(n+1)/2]$.

(e) For a double exponential population (e.g., with pdf $\frac{1}{2}\, e^{-|x-\varphi|}$), all $a(i) > 0$, $a(r) = a(r+1)$, $a(r+1) < a[(n+1)/2]$, and A decreases as population shape approaches normality.

(f) For a population with heavy tails (e.g., with pdf $C \exp[-|x-\varphi|^{1/2}]$), $a(r) = a(r+1) = 0$, and A decreases as the tails become heavier (tending to its lower limit for Cauchy distribution).

These rules should be adequate for handling most situations of practical interest.

Permutation tests and confidence intervals

Here the most general situation is considered; namely, the observations are only required to be independent and from populations with common central value φ. The problem is to obtain permutation tests for the case where $r = 1$ by use of the statistic $\hat{\varphi}$. The resulting test is first stated for all sample sizes. Then an approximation is given that can often be used when the sample size is at least moderately large.

The null hypothesis asserts that $\varphi = \varphi_0$ and the observations whose values are different from φ_0 are denoted by $y(j)$, $j = 1, ..., N \leqslant n$. Then, for the type of permutation test developed, the values of $|y(1)-\varphi_0|, ..., |y(N)-\varphi_0|$ are considered to be fixed and probability enters only with respect to the signs given these values. Since the observations yielding the $y(j)$ are independent and from symmetrical populations with common central value φ, all 2^N ways of assigning a plus or a minus sign to each of these absolute values are equally likely under the null hypothesis (see Eeden and Benard, 1957).

Establish a one-to-one correspondence between the $x(i)$ that are not equal to φ_0 and the $y(j)$, such that $x(i) = y(j)$ and $x(i)$ corresponds to $y(j)$. Then, for these $x(i)$, the value of i is determined by that of j; notationally, $i = i(j)$ for these values of i.

Let $\epsilon = (\epsilon_1, ..., \epsilon_N)$ be a vector such that each ϵ_j is either $+1$ or -1 and let $Z_1(\epsilon) \leqslant ... \leqslant Z_N(\epsilon)$ be the order statistics of $\epsilon_1|y(1)-\varphi_0|, ..., \epsilon_N|y(N)-\varphi_0|$. Consider the set of 2^N values (not necessarily unequal) obtained from

$$A(\epsilon, \varphi_0) = \sum_{j=1}^{N} a[i(j)]Z_j(\epsilon)$$

by letting ϵ take each of its 2^N values. Here $a[i(j)] = a(i)$ is the coefficient selected for $x(i)$ in the estimate $\hat{\varphi}$. One of the 2^N values of the set, corresponding to $\text{sgn}(\epsilon_j) = \text{sgn}[y(j)-\varphi_0]$ for all j, is

$$\sum_{j=1}^{N} a[i(j)][y(j)-\varphi_0] = \sum_{i=1}^{n} a(i)[x(i)-\varphi_0] = \hat{\varphi}-\varphi_0.$$

This value is used as the test statistic.

If $\hat{\varphi}-\varphi_0$ exceeds exactly $100(1-\alpha)$ percent of the 2^N values for $A(\epsilon, \varphi_0)$, the null hypothesis is rejected in favour of the alternative $\varphi > \varphi_0$; the significance level of this one-sided test is α. If $\hat{\varphi}-\varphi_0$ is less than exactly 100β percent of the values, the alternative $\varphi < \varphi_0$ is accepted at significance level β. Two-sided tests can be obtained as non-overlapping combinations of these two types of one-sided tests. Fisher's randomization test based on the observation mean (Fisher, 1949) and Nair's randomization test based on the observation median (Nair, 1940) furnish special cases of the class of permutation tests presented.

Unless N is small, direct application of these tests can be very tedious. However, if N is at least moderately large and some other conditions are satisfied, an approximation can be used. Let

$$c(\varphi_0) = (1/2) \min_{u \neq v} \Big| \, |a[i(u)][y(u)-\varphi_0]| - |a[i(v)][y(v)-\varphi_0]| \, \Big|,$$

$$\sigma^2(\varphi_0) = \sum_{j=1}^{N} a^2[i(j)][y(j)-\varphi_0]^2,$$

and suppose that $N \geqslant 25,\ 1/4N \leqslant \alpha \leqslant 1-1/4N,$

and
$$\max_j\ a^2[i(j)][y(j)-\varphi_0]^2/\sigma^2(\varphi_0) \leqslant 5/2N;$$

then, when the null hypothesis holds, on the basis of [Eden and Benard, 1957],

$$P[\hat{\varphi}-\varphi_0 < c(\varphi_0)-K_\alpha\sigma(\varphi_0)] \doteq \alpha, \quad P[\hat{\varphi}-\varphi_0 > -c(\varphi_0)+K_\alpha\sigma(\varphi_0)] \doteq \alpha;$$

here K_α is the deviate of the standardized normal distribution (zero mean and unit variance) that is exceeded with probability α. These approximate probability relations can be used to obtain one-sided and two-sided tests in the usual manner.

Confidence intervals from tests. Let $L_\alpha(\varphi_0)$ and $U_\beta\ (\varphi_0)$ be two values of $A(\epsilon, \varphi_0)$ such that exactly $100(1-\alpha)$ percent of the $A(\epsilon, \varphi_0)$ have values exceeding $L_\alpha(\varphi_0)$ and exactly 100β percent have values less than $U_\beta\ (\varphi_0)$. Also let $\varphi_L\ (\alpha)$ be the solution for φ_0 of

$$\hat{\varphi}-\varphi_0 = L_\alpha(\varphi_0)$$

while $\varphi_U(\beta)$ is the solution for φ_0 of

$$\hat{\varphi}-\varphi_0 = U_\alpha\ (\varphi_0).$$

Then,
$$P[\varphi_L\ (\alpha) < \varphi] = \alpha, \quad P[\varphi_U\ (\beta) > \varphi] = \beta, \text{ and, if } \alpha+\beta < 1,$$
$$P[\varphi_U(\beta) \leqslant \varphi \leqslant \varphi_L(\alpha)] = 1-\alpha-\beta.$$

When $N \geqslant 25$ and $1/4N \leqslant \alpha \leqslant 1-1/4N$, $\varphi_L\ (\alpha)$ can be approximately determined from

$$\hat{\varphi}-\varphi_0 = -c(\varphi_0)+K_\alpha\sigma(\varphi_0),$$

subject to the condition that the resulting value satisfies

$$\max_j\ a^2[i(j)][y(j)-\varphi_L(\alpha)]^2/\sigma^2[\varphi_L(\alpha)] \leqslant 5/2N. \qquad \ldots \quad (2)$$

Likewise, $\varphi_U(\beta)$ can be approximately determined from

$$\hat{\varphi}-\varphi_0 = c(\varphi_0)-K_\beta\sigma(\varphi_0)$$

if the resulting value satisfies

$$\max_j\ a^2[i(j)][y(j)-\varphi_U(\beta)]^2/\sigma^2[\varphi_U(\beta)] \leqslant 5/2N. \qquad \ldots \quad (3)$$

If both (2) and (3) are satisfied, the two-sided confidence interval for φ can be used on the basis of these approximate values for $\varphi_L\ (\alpha)$ and $\varphi_U\ (\beta)$.

SYMMETRY PROPERTIES

The purpose of this section is to verify that $\hat{\varphi}$ has a distribution which is symmetrical about φ when the underlying observations are independent and from symmetrical populations that have a common central value φ.

Theorem:

$$P(\hat{\varphi}-\varphi \leqslant X) = P(\hat{\varphi}-\varphi \geqslant -X)$$

for all values of X.

Proof: Let the observations be denoted by $w(1), \ldots, w(n)$. Then, since the populations are symmetrical about φ, the random variable $2\varphi-w(i)$ has the same distribution as $w(i)$. This implies that the joint distribution of the vector $2\varphi-x(n+1-r), \ldots, 2\varphi-x(r)$ is the same as that of the vector $x(r), \ldots, x(n+1-r)$. Since $a(i) = a(n+1-i)$, the variable $2\varphi-\hat{\varphi}$ has the same distribution as $\hat{\varphi}$. Thus, replacing $\hat{\varphi}$ by $2\hat{\varphi}-\varphi$,

$$P(\hat{\varphi}-\varphi \leqslant X) = P(\varphi-\hat{\varphi} \leqslant X) = P(\hat{\varphi}-\varphi \geqslant -X)$$

and the theorem is verified.

REFERENCES

FISHER, R. A. (1949): *The Design of Experiments* (5th Edition), Hafner Publishing Company, Inc., New York, 43-47.

NAIR, K. R. (1940): The median in tests by randomization. *Sankhyā*, **4**, 543-50.

SARHAN, A. E. (1955a): Estimation of the mean and standard deviation by order statistics. Part II, *Ann. Math. Stat.*, **26**, 505-11.

——— (1955b): Estimation of the mean and standard deviation by order statistics. Part III, *Ann. Math. Stat.*, **26**, 576-92.

VAN EEDEN, C. and BENARD, A. (1957): A general class of distribution-free tests for symmetry containing the tests of Wilcoxon and Fisher, I, II, III, *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen*, Series A, **60**, 382-408.

Paper received: June, 1959.

ON SOME CLASSIFICATION PROBLEMS—I

By S. JOHN

Indian Statistical Institute, Calcutta

SUMMARY. A classification procedure is described and its performance characteristics are studied. The procedure described is one which can be adopted in situations where the only information on alternative populations is contained in samples.

1. Introduction

In problems of discrimination we have an individual, known to belong to one or other of k populations and we want to assign it to its correct population. Such problems of classification often arise in several branches of science. The technique of discriminant functions which was devised by Fisher (1936) has proved to be invaluable in tackling classification problems. But the construction of the discriminant function is possible only when we know the values of the parameters characterizing the populations to be discriminated between. This raises the question as to what is to be done when such knowledge is absent. In this paper we describe a classification procedure suited to such situations and make a detailed study of the chances of error involved in this procedure.

Let us first consider the case of two populations P_1 and P_2. We shall suppose P_1 and P_2 to be multivariate normal with means $\mu^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)} \ldots, \mu_p^{(1)})$, $\mu^{(2)} = (\mu_1^{(2)}, \mu_2^{(2)}, \ldots, \mu_p^{(2)})$ respectively and a common dispersion matrix Σ. We shall, in this paper, suppose that Σ is known. We have an individual with measurements $x = (x_1, x_2, \ldots, x_p)$ and we want to assign him to P_1 or P_2. To this end we draw samples of size N_1 and N_2 from P_1 and P_2 and perform tests of significance to see whether x could have come from P_1 or P_2. Suppose the test rejects the hypothesis that x came from P_1 at a level α_1 and the hypothesis that x came from P_2 at a level α_2. We assign x to P_1 or P_2 according as $\alpha_1 \lessgtr \alpha_2$. Let $\bar{x}^{(1)}$ be the vector of sample means in the first sample and let $\bar{x}^{(2)}$ be the vector of sample means in the second sample. We may, for the tests of significance, use the criteria $(N_1/N_1+1)(\bar{x}^{(1)}-x)\Sigma^{-1}(\bar{x}^{(1)}-x)'$ and $(N_2/N_2+1)(\bar{x}^{(2)}-x)\Sigma^{-1}(\bar{x}^{(2)}-x)'$ respectively, each being distributed as a chi-square with p degrees of freedom under the respective null hypotheses. In this case our procedure amounts to assigning x to P_1 or P_2 according as

$$(N_1/N_1+1)(\bar{x}^{(1)}-x)\Sigma^{-1}(\bar{x}^{(1)}-x)' \lessgtr (N_2/N_2+1)\,(\bar{x}^{(2)}-x)\,\Sigma^{-1}(\bar{x}^{(2)}-x)' \quad \ldots \quad (1.1)$$

In the following sections we shall study in some detail the performance characteristics of such a procedure. A similar study of other classification procedures will be made in subsequent papers.

2. Probabilities of misclassification

If x really came from P_1 what is the probability of wrongly assigning it to P_2? If x really came from P_2 what is the chance of wrongly assigning it to P_1? In other words, we require

$$Pr(z_1 > z_2 | x \epsilon P_1) \text{ and } Pr(z_1 < z_2 | x \epsilon P_2) \quad \ldots \quad (2.1)$$

where

$$z_1 = (N_1/N_1+1)(\bar{x}^{(1)}-x)\, \Sigma^{-1}(\bar{x}^{(1)}-x)' \quad \ldots \quad (2.2)$$

and

$$z_2 = (N_2/N_2+1)(\bar{x}^{(2)}-x)\, \Sigma^{-1}(\bar{x}^{(2)}-x)'. \quad \ldots \quad (2.3)$$

We shall explain how to evaluate $Pr(z_1 > z_2 | x \epsilon P_1)$; $Pr(z_1 < z_2 | x \epsilon P_2)$ can be evaluated similarly.

Put

$$y^{(1)} = (N_1/N_1+1)^{\frac{1}{2}}(\bar{x}^{(1)}-x)A \quad \ldots \quad (2.4)$$

and

$$y^{(2)} = (N_2/N_2+1)^{\frac{1}{2}}(\bar{x}^{(2)}-x)A \quad \ldots \quad (2.5)$$

where A is a $(p \times p)$ matrix such that $A' \Sigma A = I$.

We can now write

$$z_1 = y^{(1)}y^{(1)'}, \quad z_2 = y^{(2)}y^{(2)'}. \quad \ldots \quad (2.6)$$

The joint distribution of $(y^{(1)}, y^{(2)})$ is multivariate normal with means $(0, \theta)$ and dispersion matrix

$$\begin{pmatrix} I & \rho I \\ \rho I & I \end{pmatrix}$$

where

$$\theta = \left(\frac{N_2}{N_2+1}\right)^{\frac{1}{2}} (\mu^{(2)}-\mu^{(1)})A \quad \ldots \quad (2.7)$$

and

$$\rho = \frac{N_1^{1/2} N_2^{1/2}}{(N_1+1)^{\frac{1}{2}}(N_2+1)^{\frac{1}{2}}}. \quad \ldots \quad (2.8)$$

We now make one more transformation and put

$$w^{(1)} = y^{(1)}C, \quad w^{(2)} = y^{(2)}C \quad \ldots \quad (2.9)$$

where C is an orthogonal matrix whose first column is

$$\left(\frac{\theta_1}{\left[\sum\limits_{i=1}^{p} \theta_i^2\right]^{\frac{1}{2}}}, \frac{\theta_2}{\left[\sum\limits_{i=1}^{p} \theta_i^2\right]^{\frac{1}{2}}}, \ldots, \frac{\theta_p}{\left[\sum\limits_{i=1}^{p} \theta_i^2\right]^{\frac{1}{2}}} \right)'.$$

In terms of the variables $w^{(1)}$ and $w^{(2)}$,

$$z_1 = w^{(1)}w^{(1)'}, \quad z_2 = w^{(2)}w^{(2)'}. \quad \ldots \quad (2.10)$$

The joint density function of $(w^{(1)}, w^{(2)})$ splits up into p factors and is, in fact,

$$(w_1^{(1)}, w_1^{(2)}) \prod_{i=2}^{p} g(w_i^{(1)}, w_i^{(2)}) \qquad \ldots \quad (2.11)$$

where $$f(a, b) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} e^{-\frac{1}{2(1-\rho^2)}[a^2-2\rho a\,(b-\delta)+(b-\delta)^2]} \qquad \ldots \quad (2.12)$$

$$g(a, b) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} e^{-\frac{1}{2(1-\rho^2)}[a^2-2\rho ab+b^2]} \qquad \ldots \quad (2.13)$$

and $$\delta = (\sum_{i=1}^{p} \theta_i^2)^{\frac{1}{2}} = \{(N_2/N_2+1)(\mu^{(2)}-\mu^{(1)})\Sigma^{-1}(\mu^{(2)}-\mu^{(1)})'\}^{\frac{1}{2}} \quad \ldots \quad (2.14)$$

δ is thus a simple multiple of the distance between P_1 and P_2.

Now put

$$u = [2(1-\rho)]^{-\frac{1}{2}}(w^{(2)}-w^{(1)}) \qquad \ldots \quad (2.15)$$

$$v = [2(1+\rho)]^{-\frac{1}{2}}(w^{(2)}+w^{(1)}). \qquad \ldots \quad (2.16)$$

Clearly $z_1 > z_2$ if and only if $uv' < 0$. Thus we get the result

$$Pr(z_1 > z_2 | x \in P_1) = Pr(uv' < 0). \qquad \ldots \quad (2.17)$$

We now address ourselves to the task of evaluating $Pr(uv' < 0)$. We observe that $u_1, u_2, \ldots, u_p$; $v_1, v_2, \ldots, v_p$ are all independently distributed normal variables. The variance of each is unity. All of them except u_1 and v_1, have got zero expectation. But

$$E(u_1) = [2(1-\rho)]^{-\frac{1}{2}}\,\delta = \eta \text{ (say)} \qquad \ldots \quad (2.18)$$

and $$E(v_1) = [2(1+\rho)]^{-\frac{1}{2}}\,\delta = \zeta \text{ (say)}. \qquad \ldots \quad (2.19)$$

First we shall find $Pr(uv' < 0|u)$ and finally get $Pr(uv' < 0)$ by finding $E[Pr(uv' < 0|u)]$.

$$Pr(uv' < 0|u) = \int_{-\infty}^{-\zeta u_1(\Sigma u_i^2)^{-\frac{1}{2}}} \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-t^2/2}\,dt$$

$$= \int_{\zeta u_1(\Sigma u_i^2)^{-\frac{1}{2}}}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}} e^{-t^2/2}\,dt. \qquad \ldots \quad (2.20)$$

We now make the substitution

$$u_1 = r\cos t, \quad \chi = (u_2^2+u_3^2+\ldots+u_p^2)^{\frac{1}{2}} = r\sin t. \qquad \ldots \quad (2.21)$$

Then

$$Pr(uv' < 0 \mid u) = \int\limits_{\zeta\cos t}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}}\, e^{-t^2/2}\, dt. \qquad \ldots \quad (2.22)$$

We observe that $Pr(uv' < 0 \mid u)$ depends on u only through t. Therefore if $\varphi(t)$ is the density function of t,

$$Pr(uv' < 0) = \int \Bigg(\int\limits_{\zeta\cos t}^{\infty} (2\pi)^{-\frac{1}{2}}\, e^{-\frac{1}{2}x^2} \Bigg)\, \varphi(t)\, dt. \qquad \ldots \quad (2.23)$$

We have thus to find the distribution of t.

3. DISTRIBUTION OF t

The joint density of u_1 and χ^2 is easily seen to be

$$2^{-(p-1)/2}\left[\Gamma\left(\frac{p-1}{2}\right)\right]^{-1} (\chi^2)^{(p-3)/2}\; e^{-\chi^2/2}\, (2\pi)^{-\frac{1}{2}}\; e^{-\frac{1}{2}(u_1-\eta)^2}. \qquad \ldots \quad (3.1)$$

Making the substitution

$$u_1 = r\cos t, \quad \chi = r\sin t \qquad \ldots \quad (3.2)$$

we find that the joint density of r and t is

$$\pi^{-\frac{1}{2}}\, 2^{-(p-2)/2}\left[\Gamma\left(\frac{p-1}{2}\right)\right]^{-1}\; e^{-\eta^2/2}\; e^{-r^2/2+\eta r\cos t}\, r^{p-1}\, \sin^{p-2} t. \qquad \ldots \quad (3.3)$$

Expanding $e^{\eta r\cos t}$ in powers of r and integrating out r we get the density function of t as

$$\varphi(t) = \pi^{-\frac{1}{2}}\left[\Gamma\left(\frac{p-1}{2}\right)\right]^{-1} e^{-\eta^2/2}\sin^{p-2} t \sum_{s=0}^{\infty} \frac{\Gamma[(p+s)/2]}{\Gamma(s+1)}\, (2\eta^2)^{s/2}\cos^s t \qquad \ldots \quad (3.4)$$

Substituting in (2.23) we get

$$Pr(uv' < 0) = \pi^{-\frac{1}{2}}\left[\Gamma\left(\frac{p-1}{2}\right)\right]^{-1} e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{\Gamma[(p+s)/2]}{\Gamma(s+1)}\, (2\eta^2)^{s/2} h_s(\zeta) \qquad \ldots \quad (3.5)$$

where

$$h_s(\zeta) = \int\limits_0^{\pi} \sin^{p-2} t\,\cos^s t\, \Bigg(\int\limits_{\zeta\cos t}^{\infty} (2\pi)^{-\frac{1}{2}}\, e^{-x^2/2}\, dx\Bigg)\, dt. \qquad \ldots \quad (3.6)$$

We shall obtain a series expansion for $h_s(\zeta)$ valid for all s.

$$h_s(\zeta) = \int_0^\pi \sin^{p-2} t \ \cos^s t \left[(2\pi)^{-\frac{1}{2}} \int_{\zeta \cos t}^{\infty} e^{-x^2/2}\, dx\right] dt$$

$$= \int_0^\pi \sin^{p-2} t \ \cos^s t \left[\tfrac{1}{2} - \int_0^{\zeta \cos t} (2\pi)^{-\frac{1}{2}} e^{-x^2/2}\, dx\right] dt \qquad \dots \quad (3.7)$$

$$= \tfrac{1}{2} \int_0^\pi \sin^{p-2} t \ \cos^s t \ dt -$$

$$- \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_{k=1}^{\infty} \frac{(-1)^k}{2^k} \frac{\zeta^{2k+1}}{k\,!} \cdot \frac{1}{2k+1} \int_0^\pi \sin^{p-2} t \ \cos^{2k+1+s} t \ dt$$

$$= \begin{cases} \tfrac{1}{2} \dfrac{\Gamma\left(\frac{p-1}{2}\right)\Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{p+s}{2}\right)} . \ \text{(if } s \text{ is even)} \\[2ex] -(2\pi)^{-\frac{1}{2}} \displaystyle\sum_{k=0}^{\infty} \left(-\tfrac{1}{2}\right)^k \frac{\zeta^{2k+1}}{k!(2k+1)} \cdot \frac{\Gamma\left(\frac{p-1}{2}\,\Gamma\right), \left(\frac{2k+s+2}{2}\right)}{\Gamma\left(\frac{2k+p+s+1}{2}\right)} \end{cases} \qquad \dots \quad (3.8)$$

(if s is odd).

It will not be difficult for the reader to verify that the term by term integration involved is justified.

4. Bounds for the probability

In (3.8) we have a very simple expression for $h_s(\zeta)$ if s is even and an infinite series if s is odd. We shall now determine upper and lower bounds for $h_s(\zeta)$ for s odd. These will in their turn determine upper and lower bounds for $Pr(uv' < 0)$.

If s is odd, we have from (3.7)

$$h_s(\zeta) = -\int_0^\pi \sin^{p-2} t \cos^s t \left(\int_0^{\zeta \cos t} (2\pi)^{-\frac{1}{2}} \ e^{-u^2/2} \ du \right)$$

$$= -2 \int_0^1 (1-x^2)^{\frac{p-3}{2}} x^s \left(\int_0^{\zeta x} (2\pi)^{-\frac{1}{2}} \, e^{-u^2/2} \, du \right) dx. \qquad \dots \quad (4.1)$$

Observe that $\int_0^{\zeta x} (2\pi)^{-\frac{1}{2}}\, e^{-u^2/2}\, du$ is a concave function of x. Hence we have

$$h_s(\zeta) \geqslant - \frac{\Gamma\left(\frac{p-1}{2}\right)\Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{p+s}{2}\right)} \int_0^{\zeta a(s)} (2\pi)^{-\frac{1}{2}}\, e^{-u^2/2}\, du \qquad \ldots \quad (4.2)$$

where

$$a(s) = \frac{\Gamma\left(\frac{s+2}{2}\right).\,\Gamma\left(\frac{p+s}{2}\right)}{\Gamma\left(\frac{s+1}{2}\right).\,\Gamma\left(\frac{p+s+1}{2}\right)} \qquad \ldots \quad (4.3)$$

Further it is obvious that, for s odd

$$h_s(\zeta) \leqslant 0. \qquad \ldots \quad (4.4)$$

Using (3.8) and (4.2) we obtain from (3.5) the following bounds for $Pr(uv' < 0)$

$$\pi^{-\frac{1}{2}}\, e^{-\eta^2/2} \left[\tfrac{1}{2} \sum_{s=0}^{\infty} \frac{\Gamma(s+\frac{1}{2})}{\Gamma(2s+1)} (2\eta^2)^s - \sum_{s=0}^{\infty} \frac{\Gamma(s+1)}{\Gamma(2s+2)} \left(\int_0^{\zeta a(2s+1)} (2\pi)^{-\frac{1}{2}}\, e^{-u^2/2} du \right) (2\eta^2)^{s+\frac{1}{2}} \right]$$

$$\leqslant Pr(uv' < 0) \leqslant \tfrac{1}{2}\pi^{-\frac{1}{2}}\, e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{\Gamma(s+\frac{1}{2})}{\Gamma(2s+1)} (2\eta^2)^s. \qquad \ldots \quad (4.5)$$

These bounds can be written in a simpler form using the duplication formula for the gamma function

$$\pi^{\frac{1}{2}}\Gamma(2x) = 2^{2x-1}\Gamma(x).\ \Gamma(x+\tfrac{1}{2})$$

$$\tfrac{1}{2} - e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{(\eta^2/2)^{s+\frac{1}{2}}}{\Gamma(s+3/2)} \left(\int_0^{\zeta a(2s+1)} (2\pi)^{-\frac{1}{2}} e^{-u^2/2}\, du \right) \leqslant Pr(uv' < 0) \leqslant \tfrac{1}{2}. \quad \ldots \quad (4.6)$$

These bounds are easier to calculate than the exact expression for $P(uv' < 0)$ itself.

A slightly more simple, but less sharp, lower bound for $Pr(uv' < 0)$ is obtained if we use the fact that, for odd s

$$h_s(\zeta) \leqslant - \frac{\Gamma\left(\frac{p-1}{2}\right).\ \Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{p+s}{2}\right)} \int_0^\zeta (2\pi)^{-\frac{1}{2}}\, e^{-u^2/2}\, du$$

$$\tfrac{1}{2} - e^{-\eta^2/2}\left(\int_0^\zeta (2\pi)^{-\frac{1}{2}}\, e^{-u^2/2}\, du\right)\left[\sum_{s=0}^{\infty} \frac{(\eta^2/2)^{s+\frac{1}{2}}}{\Gamma(s+3/2)}\right] \leqslant Pr(uv' < 0) \leqslant \tfrac{1}{2} \quad \ldots \text{(4.7}$$

It is possible to give yet another lower bound. (I am indebted to Dr. C. R. Rao for pointing out this). This is obtained by noting that

$$h_{2s+1}(\zeta) \leqslant |h_{2s+1}(\zeta)| \leqslant h_{2s}(\zeta).$$

Hence we get after some simplification,

$$\tfrac{1}{2} - e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{(\eta^2/2)^{s+\frac{1}{2}}}{s\,!} \cdot \frac{\Gamma\left(s + \frac{p+1}{2}\right)}{\Gamma(s+p/2)} \leqslant Pr(uv' < 0). \qquad \ldots \quad (4.8)$$

If we separate even and odd terms, the exact formula for $Pr(uv' < 0)$ itself can be expressed more simply as

$$Pr(uv' < 0) = \tfrac{1}{2} + \pi^{-\frac{1}{2}}\left[\Gamma\left(\frac{p-1}{2}\right)\right]^{-1} e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{\Gamma\left(s + \frac{p+1}{2}\right)}{(2s+1)\,!} (2\eta^2)^{s+\frac{1}{2}}\, h_{2s+1}(\zeta) \qquad \ldots \quad (4.9)$$

5. CLASSIFICATION WHEN THERE ARE MORE THAN TWO POPULATIONS

Generalizing the procedure discussed above, we may compute

$$z_i = (N_i/N_i+1)(\bar{x}^{(i)} - x)\Sigma^{-1}(\bar{x}^{(i)} - x)' \quad (i = 1, 2, \ldots, k) \qquad \ldots \quad (5.1)$$

and assign x to P_j if $z_j < z_i$ $(i = 1, 2, \ldots, k,\ i \neq j)$. Here $\bar{x}^{(i)}$ is the vector of means computed from a sample of size N_i taken from P_i and Σ is the common dispersion matrix assumed to be known. Let us consider the probability of incorrect classification involved in the above procedure. Assume that x really belongs to P_1. We shall give an upper bound to the probability of incorrect classification. Similar upper bounds can be given for the probabilities of misclassification of individuals from $P_2, P_3, \ldots, P_k$.

Let E_i $(i = 2, 3, \ldots, k)$ be the event that $z_i < z_1$. Then the chance of misclassifying x is $Pr\,(\bigcup_{i=2}^{k} E_i)$.

Now, $$Pr\,(\bigcup_{i=2}^{k} E_i) \leqslant \sum_{i=2}^{k} Pr\,(E_i)\,. \quad \text{(Uspensky, 1937).} \qquad \ldots \quad (5.2)$$

But $Pr(E_i)$ $(i = 2, \ldots, k)$ is the probability of misclassification in the situation where there are only two populations P_1 and P_i. We have already explained how this probability can be computed.

REFERENCES

FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179.

USPENSKY, J. V. (1937): *Introduction to Mathematical Theory of Probability*, 30, McGraw Hill Book Company Inc., New York and London.

Paper received : November, 1959.

ON SOME CLASSIFICATION STATISTICS

By S. JOHN

Indian Statistical Institute, Calcutta

SUMMARY. Various authors have suggested various statistics for use in classification problems. To set up the classification procedure and to study the performance characteristics of procedures thus set up, it is necessary to know the distributions of the statistics used. This paper gives the exact distributions of several classification statistics.

1. Introduction

Research workers are often faced with the problem of assigning an individual, known to belong to one or other of two classes, to its correct class. They do this by observing various characteristics of the individual. Some characteristics whose ranges for the two classes are completely disjoint would have been ideal for classification purposes. But it is generally the case that there is a certain range of values where individuals from both the classes may appear. This prevents us from being always correct. But if we know the distributions of the characteristics in both the classes, we can devise a procedure in which the chances of misclassification are least. This will be achieved if we use the likelihood ratio as a criterion. The well-known procedure of using the discriminant function is equivalent to the one using the likelihood ratio provided the distributions of the characteristics are in both the populations, multivariate normal with identical dispersion matrices. If we denote the characteristics observed by $x = (x_1, ..., x_p)$ the common dispersion matrix by Σ and the mean vectors in the two populations P_1 and P_2 by $\mu^{(1)}$ and $\mu^{(2)}$, the discriminant function is

$$D(x) = (\mu^{(2)} - \mu^{(1)})\Sigma^{-1}x'. \qquad \ldots \quad (1.1)$$

Individuals with small $D(x)$ values are assigned to P_1 and those with large $D(x)$ values are assigned to P_2.

If the distributions of x in P_1 and P_2 are unknown, we are unable to use this procedure. To meet this situation, Wald (1944) suggested that we could use the function $U(x)$ obtained by substituting in $D(x)$ sample estimates of $\mu^{(1)}$, $\mu^{(2)}$ and Σ. To set up the classification procedure and to evaluate the chances of misclassification involved in adopting the procedure thus set up, we require the distributions of $U(x)$ for x's from P_1 and P_2. Wald tackled this problem in 1944. He exhibited $U(x)$ as a function of three statistics m_1, m_2, m_3 whose joint distribution he expressed as a product of three factors, the first an exponential factor, the second a product of gamma and beta functions and the third an expectation of a power of a random determinant. In 1951, Anderson evaluated this expectation for the special case where $\mu^{(1)}$ and $\mu^{(2)}$ are parallel vectors. Later Sitgreaves (1952) gave an alternate derivation of the joint distribution of m_1, m_2 and m_3, again for the same special case as considered by Anderson. Still it was only the joint distribution of m_1, m_2 and m_3 and not the distribution of Wald's classification statistic itself. Harter (1951) derived the exact distribution for the univariate case under the assumption that at least one of the populations has zero mean. For the multivariate case he showed that an approximation could be derived provided the mean vector of one of the populations is zero and the dispersion matrix is estimated with a large even degree of freedom.

In a previous paper (John, 1959), the author obtained the exact distribution of Wald's classification statistic for the case where the dispersion matrix is known. Unfortunately, in that paper the assumption was made that there was no loss of generality in assuming $\mu^{(1)} = 0$. In this paper, we remove this restriction.

Since Wald made his suggestion other authors have come forward with other procedures. Of these, the one suggested by Rao (1954) was specially constructed so as to be most powerful in discriminating between alternatives which were close to each other. But in no case was the necessary distribution available. We obtain these distributions. Some remarks on the relative merits of these statistics are also included.

2. NOTATION

We shall denote the two alternate populations by P_1 and P_2. It is assumed that the classification procedure is based on p characteristics $x_1, x_2, \ldots, x_p$. In what follows, $x = (x_1, \ldots, x_p)$ will be supposed to follow the multivariate normal law in both P_1 and P_2. We shall assume that these two distributions have the same dispersion matrix Σ but different mean vectors, $\mu^{(1)} = (\mu_1^{(1)}, \ldots, \mu_p^{(1)})$ for P_1 and $\mu^{(2)} = (\mu_1^{(2)}, \ldots, \mu_p^{(2)})$ for P_2.

Estimates of $\mu^{(1)}$ and $\mu^{(2)}$ are required later. (We are, in this paper, assuming that Σ is known). If $x_{i\alpha}^{(1)}$ ($\alpha = 1, 2, \ldots, N_1$; $i = 1, 2, \ldots, p$) is random sample of N_1 individuals from P_1 and $x_{i\alpha}^{(2)}$ ($\alpha = 1, \ldots, N_2$; $i = 1, 2, \ldots, p$) is a sample of N_2 individuals from P_2,

$$\bar{x}^{(1)} = (\bar{x}_1^{(1)}, \bar{x}_2^{(1)}, \ldots, \bar{x}_p^{(1)})$$

and

$$\bar{x}^{(2)} = (\bar{x}_1^{(2)}, \bar{x}_2^{(2)}, \ldots, \bar{x}_p^{(2)})$$

where

$$\bar{x}_i^{(1)} = \frac{\sum\limits_{\alpha=1}^{N_1} x_{i\alpha}^{(1)}}{N_1} \quad \text{and} \quad \bar{x}_i^{(2)} = \frac{\sum\limits_{\alpha=1}^{N_2} x_{i\alpha}^{(2)}}{N_2}$$

can be used as estimates of $\mu^{(1)}$ and $\mu^{(2)}$ respectively.

3. WALD'S CLASSIFICATION STATISTIC

In the notation developed above, the statistic suggested by Wald is

$$U(x) = (\bar{x}^{(2)} - \bar{x}^{(1)})\Sigma^{-1}x'. \quad \ldots \quad (3.1)$$

Using this statistic we may set up a classification procedure as follows : Determine a constant d_1 such that

$$Pr\,(U(x) < d_1 \,|\, x \in P_1) = \alpha_1 \quad \ldots \quad (3.2)$$

and a constant d_2 such that

$$Pr\,(U(x) < d_2 \,|\, x \in P_2) = \alpha_2. \quad \ldots \quad (3.3)$$

What α_1 and α_2 are—will be made clear below. Assign the individual to P_1 or P_2 according as

$$U(x) \begin{array}{l} < \min(d_1, d_2) \\ \geqslant \max(d_1, d_2) \end{array} \quad \ldots \quad (3.4)$$

If this classification procedure is adopted, the chance of misclassifying an individual from P_1 will not exceed α_1; so also the chance of misclassifying an individual from P_2 will not exceed α_2.

If $d_1 < d_2$ and $d_1 \leqslant U(x) \leqslant d_2$, or if $d_2 < d_1$ and $d_2 \leqslant U(x) \leqslant d_1$, we will not classify the individual as belonging to P_1 or P_2. In the first case the individuals who are unclassified are those about whom there is a suspicion that they came neither from P_1 nor from P_2 but from some other population. In the second case the individuals who are unclassified are those who are claimed by both P_1 and P_2.

If it is preferred not to have a region of indeterminacy, we may adopt the procedure of classifying the individual as belonging to P_1 or P_2 according as

$$U(x) \underset{\geqslant}{<} U(\tfrac{1}{2}[\bar{x}^{(1)} + \bar{x}^{(2)}]). \quad \ldots \quad (3.5)$$

The chances of errors of the two kinds involved in this procedure can be assessed with the help of the distributions which we obtain below.

The distribution of Wald's classification statistic. The distribution of the statistic $(\bar{x}^{(2)} - \bar{x}^{(1)})\ \Sigma^{-1}x'$ is required. We shall obtain the distribution of this statistic under the assumption that x came from P_1 : its distribution for x's from P_2 can be obtained similarly.

Rather than $(\bar{x}^{(2)} - \bar{x}^{(1)})\Sigma^{-1}x'$ we prefer to consider the statistic

$$T = 2\,\frac{(N_1N_2)^{\frac{1}{2}}}{(N_1+N_2)^{\frac{1}{2}}}\,(\bar{x}^{(2)} - \bar{x}^{(1)})\ \Sigma^{-1}x'. \quad \ldots \quad (3.6)$$

Put

$$u = xA,\ v = \frac{(N_1N_2)^{\frac{1}{2}}}{(N_1+N_2)^{\frac{1}{2}}}\ (\bar{x}^{(2)}-\bar{x}^{(1)})A \qquad \ldots \quad (3.7)$$

where A is a $(p\times p)$ matrix such that $A'\Sigma\, A = I$.

Then

$$T = 2vu' = 2\sum_{i=1}^{p} u_i\, v_i = \tfrac{1}{2}\sum_{i=1}^{p}(v_i+u_i)^2-\tfrac{1}{2}\sum_{i=1}^{p}(v_i-u_i)^2 = t-w \text{ (say)}. \qquad \ldots \quad (3.8)$$

t and w have independent noncentral chi-square distributions, in each case with degree of freedom p but with noncentralities given respectively by

$$\lambda_1 = \tfrac{1}{4}\, c\, \Sigma^{-1}c' \quad \text{and} \quad \lambda_2 = \tfrac{1}{4}\, d\, \Sigma^{-1}d' \qquad \ldots \quad (3.9)$$

where

$$c = \mu^{(1)}+(N_1^{-1}+N_2^{-1})^{-\frac{1}{2}}(\mu^{(2)}-\mu^{(1)}) \qquad \ldots \quad (3.10)$$

and

$$d = \mu^{(1)}-(N_1^{-1}+N_2^{-1})^{-\frac{1}{2}}(\mu^{(2)}-\mu^{(1)}) \qquad \ldots \quad (3.11)$$

The joint density of t and w is

$$f(t, w)$$

$$= 2^{-p}\ e^{-\lambda_1-\lambda_2}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{2^{-r-s}\,\lambda_1^r\,\lambda_2^s}{r\,!\ s\,!\ \Gamma\left(\frac{p+2r}{2}\right)\Gamma\left(\frac{p+2s}{2}\right)}\ t^{(p+2r-2)/2}\ w^{(p+2s-2)/2}\ e^{-(t+w)/2}.$$

Hence the density function of T is given by ... (3.12)

$$g(T)dT = \int_{T<t-w<T+dT}\!\!\cdots\!\int f(t, w)dtdw. \qquad \ldots \quad (3.13)$$

On simplification this gives

$$g(T) = 2^{-p}\ e^{-\lambda_1-\lambda_2}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{2^{-r-s}\,\lambda_1^r\,\lambda_2^s}{r\,!\ s\,!\ \Gamma\left(\frac{p+2r}{2}\right)}\ T^{(r+s+p-2)/2}\ W_{l,m}\,(T)(T>0) \qquad \ldots \quad (3.14)$$

where

$$l = \frac{r-s}{2} \quad m = \frac{r+s+p-1}{2} \qquad \ldots \quad (3.15)$$

and where $W_{l,m}\,(T)$ is Whittaker's confluent hypergeometric function defined by

$$W_{l,m}(x) = \frac{x^{m+\frac{1}{2}}}{\Gamma(m-l+\frac{1}{2})}e^{-x/2}\int^{\infty} e^{-tx}\ t^{m-l-\frac{1}{2}}\,(1+t)^{m+l-\frac{1}{2}}\,dt. \qquad \ldots \quad (3.16)$$

If $T<0$,

$$g(T) = 2^{-p}\ e^{-\lambda_1-\lambda_2}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{2^{-r-s}\,\lambda_1^r\lambda_2^s}{r!\ \ s!\ \Gamma\left(\frac{p+2s}{2}\right)}\ (-T)^{(r+s+p-2)/2}\ W_{-l,m}\,(-T) \qquad \ldots \quad (3.17)$$

4. ANDERSON'S CLASSIFICATION STATISTIC

In this section we consider a statistic suggested by Anderson for classification purposes. In our notation this statistic is

$$Q = (\bar{x}^{(2)}-\bar{x}^{(1)})\Sigma^{-1}x'-\tfrac{1}{2}(\bar{x}^{(2)}-\bar{x}^{(1)})\Sigma^{-1}(\bar{x}^{(2)}+\bar{x}^{(1)})'. \qquad \ldots \ (4.1)$$

The steps involved in setting up the classification procedure are much the same as for Wald's classification statistic. We determine two constants c_1 and c_2 such that

$$Pr(Q > c_1 \,|\, x \in P_1) = \alpha_1 \qquad \ldots \ (4.2)$$

and

$$Pr(Q < c_2 \,|\, x \in P_2) = \alpha_2 \qquad \ldots \ (4.3)$$

α_1 and α_2 being respectively the levels at which the two kinds of error are to be controlled. Individuals are assigned to P_1 or P_2 according as

$$Q \begin{matrix} < \min\,(c_1, c_2) \\ \geqslant \max\,(c_1, c_2). \end{matrix} \qquad \ldots \ (4.4)$$

If this procedure is adopted the probability of wrongly assigning to P_2 an individual from P_1 will not exceed α_1; so also the probability of wrongly assigning to P_1 an individual from P_2 will not exceed α_2.

It is to be noted that we are abstaining from classifying individuals for whom

$$\min\,(c_1, c_2) \leqslant Q(x) \leqslant \max\,(c_1, c_2). \qquad \ldots \ (4.5)$$

These are again individuals who are either claimed by both P_1 and P_2 or individuals about whom there is a suspicion that they came neither from P_1 nor from P_2 but from some other population.

If it is desired not to have a region of indecision the following classification procedure may be followed. Assign individuals to P_1 or P_2 according as

$$Q \lesseqgtr 0. \qquad \ldots \ (4.6)$$

It is not difficult to see that this procedure is equivalent to assigning individuals to P_1 or P_2 according as

$$U(x) \lesseqgtr U(\tfrac{1}{2}[\bar{x}^{(1)}+\bar{x}^{(2)}]). \qquad \ldots \ (4.7)$$

In whatever way the classification procedure is set up, if it is based on Q, the chances of errors of the two kinds involved in such a procedure can be evaluated using the distribution of Q obtained below. The author has given elsewhere simple approximations as well as exact expressions for these probabilities.

Distribution of Anderson's classification statistic. We shall now obtain the distribution of Anderson's classification statistic for x's from P_1; its distribution for x's from P_2 be obtained similarly.

Since the steps involved in arriving at these distributions are similar to those for Wald's classification statistic, only the final expressions will be given.

$$z = 4N_1N_2(N_1+N_2)^{-\frac{1}{2}}(N_1+N_2+4N_1N_2)^{-\frac{1}{2}}Q \qquad \ldots \quad (4.8)$$

has the density function

$$f_1(z) = \begin{cases} e^{-z(a-b)/4} \sum\limits_{r=0}^{\infty} \sum\limits_{s=0}^{\infty} c_{rs} \dfrac{a^{\frac{1}{2}(p+2r)} b^{\frac{1}{2}(p+2s)}}{\Gamma\left(\dfrac{p+2r}{2}\right)} z^{\frac{1}{2}(p+r+s)-1} W_{l,m}(cz) \; (z > 0) \\[2em] e^{-z(a-b)/4} \sum\limits_{r=0}^{\infty} \sum\limits_{s=0}^{\infty} c_{rs} \dfrac{a^{\frac{1}{2}(p+2r)} b^{\frac{1}{2}(p+2s)}}{\Gamma\left(\dfrac{p+2s}{2}\right)} (-z)^{\frac{1}{2}(p+r+s)-1} W_{-l,m}(-cz) \; (z < 0) \end{cases} \qquad \ldots \quad (4.9)$$

where

$$a = (1+\rho)^{-1}, \quad b = (1-\rho)^{-1}, \quad c = (a+b)/2 \qquad \ldots \quad (4.10)$$

$$\rho = (N_2-N_1)(N_1+N_2)^{-\frac{1}{2}}[4N_1N_2+N_1+N_2]^{-\frac{1}{2}} \qquad \ldots \quad (4.11)$$

$$c_{rs} = e^{-\lambda_1-\lambda_2} \frac{\lambda_1^r \lambda_2^s}{r!s!} (4c)^{-(p+r+s)/2} \qquad \ldots \quad (4.12)$$

$$\lambda_1 = \tfrac{1}{4}(1+\rho)^{-1}N_1N_2[(N_1+N_2)^{-\frac{1}{2}}-(N_1+N_2+4N_1N_2)^{-\frac{1}{2}}]^2\delta^2 \qquad \ldots \quad (4.13)$$

$$\lambda_2 = \tfrac{1}{4}(1-\rho)^{-1}N_1N_2[(N_1+N_2)^{-\frac{1}{2}}+(N_1+N_2+4N_1N_2)^{-\frac{1}{2}}]^2\delta^2 \qquad \ldots \quad (4.14)$$

$$\delta^2 = (\mu^{(2)}-\mu^{(1)})\Sigma^{-1}(\mu^{(2)}-\mu^{(1)})'. \qquad \ldots \quad (4.15)$$

l and m have the same meanings as before.

It is easy to see that if $N_1 = N_2$, z has the same distribution as T with λ_1 and λ_2 given by equations (4.13) and (4.14).

5. RAO'S STATISTIC

Rao (1954) proposes the following statistic for classification purposes

$$R = \frac{a_1^2-a_2^2}{b_2^2} U\Sigma^{-1}U' + \frac{2(a_1-a_2)}{b_1b_2} U\Sigma^{-1}T' \qquad \ldots \quad (5.1)$$

where

$$a_1 = N_2(N_1+N_2)^{-1},$$
$$a_2 = -N_1(N_1+N_2)^{-1}$$
$$b_1 = N_1^{-1}+N_2^{-1},$$
$$b_2 = 1+(N_1+N_2)^{-1}$$
$$U = x-(N_1\bar{x}^{(1)}+N_2\bar{x}^{(2)})(N_1+N_2)^{-1}$$
$$T = \bar{x}^{(1)}-\bar{x}^{(2)}.$$

When $N_1 = N_2$, this is equivalent to Anderson's statistic.

For later use we define

$$d = 2b_1^{-1}\, b_2(a_1+a_2)^{-1}. \qquad \dots \quad (5.2)$$

Assuming that x came from P_1,

$$R_c = \frac{b_2^2}{a_1^2-a_2^2}\,[1+(N_1+N_2)^{-1}]^{-\frac{1}{2}}[1+(N_1+N_2)^{-1}+b_1d^2]^{-\frac{1}{2}}R \qquad \dots \quad (5.3)$$

has the same distribution as z, with

$$\begin{aligned}\lambda_1 = \tfrac{1}{4}(1+\rho)^{-1}[N_2(N_1+N_2)^{-\frac{1}{2}}(N_1+N_2+1)^{-\frac{1}{2}}+ \\ +\{1+(N_1+N_2)^{-1}+b_1d^2\}^{-\frac{1}{2}}(a_1+d)]^2\delta^2\end{aligned} \qquad \dots \quad (5.4)$$

$$\begin{aligned}\lambda_2 = \tfrac{1}{4}(1-\rho)^{-1}[N_2(N_1+N_2)^{-\frac{1}{2}}(N_1+N_2+1)^{-\frac{1}{2}}- \\ -\{1+(N_1+N_2)^{-1}+b_1d^2\}^{-\frac{1}{2}}(a_1+d)]^2\delta^2\end{aligned} \qquad \dots \quad (5.5)$$

$$\rho = [1+(N_1+N_2)^{-1}]^{\frac{1}{2}}[1+(N_1+N_2)^{-1}+b_1d^2]^{-\frac{1}{2}}. \qquad \dots \quad (5.6)$$

The changes to be made to obtain the distribution of this statistic for x's from P_2 are obvious.

The distributions we have obtained are found to contain certain parameters. The distributions of Anderson's classification statistic and Rao's classification statistic involve one parameter each. In either case it is the distance between the two populations P_1 and P_2. On the other hand the distribution of Wald's classification statistic contains two parameters. They are not simply related to the distance between P_1 and P_2. This feature makes Wald's classification statistic less attractive than either Anderson's statistic or Rao's statistic.

That the classification procedure using Wald's statistic is not invariant under changes of origin is another point on its debit side.

Since the distributions involve parameters, how shall we find constants c_1, c_2, d_1, d_2 etc? Consider Anderson's statistic and Rao's statistic. The parameter appearing in the distributions of either of these statistics is the distance δ between the two populations P_1 and P_2. We may require that if $\delta \geqslant \delta_0$ the probability of an error of the first kind, (i.e. of wrongly classifying an individual from P_1 as belonging to P_2) shall not exceed α_1 and also that the probability of an error of the second kind (i.e. of wrongly classifying an individual from P_2 as belonging to P_1) shall not exceed α_2. This requirement will be satisfied if we determine the constants c_1, c_2 etc., assuming that $\delta = \delta_0$. For a classification procedure thus set up, the probability of errors of the two kinds for other values of δ can be evaluated with the help of the distributions obtained in this paper.

A discussion of further problems of a similar nature will be included in a later paper.

I wish to thank Dr. C. R. Rao for his careful perusal of this paper.

REFERENCES

ANDERSON, T. W. (1951) : Classification by multivariate analysis. *Psychometrika*, **16**, 31.

HARTER, H. L. (1951) : On the distribibution of Wald's classification statistic. *Ann. Math. Stat.*, **22**, 58.

JOHN, S. (1959) : The distribution of Wald's classification statistic when the dispersion matrix is known. *Sankhyā*, **21**, 371.

RAO, C. R. (1954) : On a general theory of discrimination when the information on alternate hypotheses is based on samples. *Ann. Math. Stat.*, **25**, 651.

SITGREAVES, R. (1952) : On the distribution of two random matrices used in classification procedures. *Ann. Math. Stat.*, **23**, 263.

WALD, A. (1944) : On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Stat.*, **15**, 145.

Paper received : November, 1959.

MULTIVARIATE ANALYSIS : AN INDISPENSABLE STATISTICAL AID IN APPLIED RESEARCH*

By C. RADHAKRISHNA RAO
Indian Statistical Institute, Calcutta

1. INTRODUCTION

Some of our past presidents had selected for their addresses topics of philosophical interest or had reviewed recent trends in statistics. A few presented current statistical theories of inference, while others explored special techniques in fields of application. I am inclined to fall in line with the last group, although with one difference. Instead of choosing a particular field of application I propose to consider that body of statistical methods generally known as multivariate analysis and discuss its usefulness in applied research in a variety of fields.

Research workers are often satisfied with univariate techniques in the reduction and interpretation of their data. Multivariate methods, though more appropriate in certain situations, are not adopted mainly because of the complexity of the procedures involved, which make the research workers dependent on specialists who can understand the situation and prescribe, with some success, suitable formulae, and also on computers capable of receiving complicated instructions for numerical work.

Statisticians themselves are not unanimous in their views about the role of multivariate analysis. Fortunately, however, the controversies have not been of the traditional type we are used to in statistics, leading to the creation of rival schools of thought.

A few statisticians are optimistic and think that multivariate analysis 'has reached a fairly stable position theoretically and a considerable service might be done if further discussion were attempted of its use in different fields'.[1] Some doubt the real superiority of multivariate methods and consider the examination of individual measurements by simpler univariate methods 'as an *essential step*, which can be supplemented by more elaborate multivariate analysis when this appears feasible and necessary'.[2] A few others are critical of particular applications and regard some papers on multivariate analysis as providing 'new outlets for mathematical theory without materially assisting scientific research' and suggest 'the need to examine carefully their relevance to the interpretation of experimental and observational data'.[3] We also have the critical remark addressed to the statisticians that 'we go multivariate when we do not understand the practical situation'.[4]

---

* Presidential Address at the fortyseventh session of the Indian Science Congress, Section of Statistics, January 1960, Bombay; published in the *Proceedings of the 47th Session, Indian Science Congress*, Part II.

[1] From private correspondence with Professor M. S. Bartlett.

[2] Yates and Healy (1951), *Nature*, 168, p. 1116.

[3] Finney (1956).

[4] From an after dinner talk by Professor J. Tukey.

The subject is not, therefore, without those controversial elements which are perhaps necessary for the proper development and refinement of techniques.

It must be remembered that a tool, although it may figure prominently for some time in literature, will condemn itself if its application is subject to severe limitations or if its practice misleads the research workers. Introduced and developed by pioneers like Galton, Pearson, Fisher and Mahalanobis for application in biometric studies and by Spearman and Hotelling in psychometric studies, multivariate analysis has now evolved itself as an essential scientific aid in applied research.

2. Multivariate analysis as techniques of prediction

In order to emphasize the practical importance of multivariate analysis, I would like to introduce the subject as essentially a body of techniques to be used for predicting one or more what we may call criterion variables (c.v.'s) with the help of others known as predictor variables (p.v.'s).

The criterion variables may have nominal, ordinal or interval scale of measurement and they may be observable or hypothetical, while the predictor variables may have nominal or interval scale (see footnote 2).

The general theory of regression subsumes, as a rule, all prediction problems but differences exist in the computational procedures for different situations distinguishable by the scale types of criterion and predictor variables. Chart 1 shows the different possible combinations and, wherever available, the statistical terms used to describe the computational procedures.[1]

CHART 1. MULTIVARIATE ANALYSIS AS PROCEDURES FOR PREDICTION

scale of predictor variable	scale of criterion variable[2]			
	nominal	ordinal	interval	
			observable	hypothetical
nominal	contigency tables	?	least square theory	latent structure analysis
interval	discriminant function	?	multiple regression	genetic selection, factor analysis

---

[1] Lubin (1950) considered the two possibilities, quantitative and qualitative for each of predictor and criterion variables.

[2] Measurements are generally classified into four distinct scales. Numbers, names or other symbols used simply to classify an object, person or characteristic constitute a *nominal* or a classificatory scale. The only relation that can exist between elements on a nominal scale is *equivalence* by which the various elements can only be identified as belonging to different groups. If besides equivalence the relation $>$ holds between the identifying symbols, then a rank ordering is possible in which case the measurement is said to be on an *ordinal* or ranking scale. If between measurements on an ordinal scale comparable distances can be defined we have an *interval* scale. An interval scale further having a true zero point for its origin is called a *ratio* scale. For example if my friend is described as (i) South Indian, (ii) Under Secretary, (iii) four inches taller (than I am), (iv) weighing 98 lbs., and (v) having 12 children, then we have 5 measurements on him of which (i) is on a nominal scale, (ii) on an ordinal scale (in government hierarchy), (iii) on an interval scale, and (iv) and (v) on a ratio scale. A measurement on a ratio scale also belongs to an interval scale and the statistical procedures described in this paper are same for both.

For the combinations where question marks are placed there appear to be no standard statistical methods, although such situations are sometimes met with in practice (see section 7 of this paper for a discussion of the ordinal-interval combination).

It is not suggested that other possibilities do not exist, such as using mixtures of nominal and interval scale variables or even using an ordinal scale variable for predicting a c.v. The scheme indicated in Chart 1 is adopted only as a suitable framework for discussing some of the different procedures available in the literature. It must be also mentioned that for a given combination of p.v. and c.v., a particular method indicated is applicable only in some special situations.

I shall consider the various methods mentioned in Chart I one by one, briefly review the present position and make some suggestions for further refinement of the method or enlarging its scope. Two of the methods, contingency tables and least square theory, are well known and need no further discussion in the present context.

3. MULTIPLE REGRESSION

(C.V.'S AND P.V.'S OBSERVABLE ON INTERVAL SCALE)

The theoretical development of the subject of multivariate analysis originated with the concept of regression, introduced and used by Francis Galton (1889) in his famous study of natural inheritance. Establishing a linear relationship between the average stature of sons and the stature of the Mid-Parents[1] he came to the following conclusions,[2] which can be equally true for many other characteristics. 'Sons do not on the average resemble their Mid-Parents except when they are mediocre... The more bountifully the parent is gifted by Nature, the more rare will be his good fortune if he begets a son who is as richly endowed as himself... although it is true that children of a gifted pair are much more likely to be gifted than children of a mediocre pair... The ablest of all the children of a few gifted pairs is not likely to be as gifted as the ablest of all children of a very great many mediocre pairs. The number of individuals who are nearly mediocre is so preponderant, that an exceptional man is more frequently found to be the exceptional son of mediocre parents than the average son of very exceptional parents.'

Karl Pearson and Udny Yule, and R. A. Fisher (who provided an exact statistical theory) extended the concept of regression for predicting a variable utilising a number of other variables. Some of the purposes for which this method is applied are considered in this section.

Prediction in an individual case. An early example in India of the use of multiple regression is in the prediction of monsoon rainfall well in advance of its onset, which is very important for an agricultural country like ours. Gilbert Walker (1914) was able to establish a good correlation between total rainfall in India with preceding weather conditions outside India. The regression equation given for this purpose in 1910 was

---

[1] This is the average value of stature of the father and the transmuted stature of the mother. Galton obtained the transmuted values (or the male equivalent) by multiplying the female measurements by the factor 13/12.

[2] The conclusions are the consequences of the regression coefficient of son's height on father's being less than unity.

$$\begin{aligned}\text{Monsoon rainfall in India} = & -0.20\ \text{(snowfall accumulation)}\\ & -0.29\ \text{(Mauritius pressure)}\\ & +0.28\ \text{(South American pressure)}\\ & -0.12\ \text{(Zanzibar rainfall).}\end{aligned}$$

which, as you may observe, has some baffling predictor variables. The real test of a regression equation is its performance, i.e. 'waiting and seeing its behaviour.' Or alternatively, if we have a long series of observations, say, over time, on all the variables, we may construct a regression formula using the earlier figures, and examine, in the rest of the series, the difference between observed values of the criterion variable and the predicted values from the regression function. When the agreement is good in some sense, a single formula may be worked out from the whole series. If not, we may have to look for other variables which may be changing over time and affecting the series of figures and whose inclusion might improve prediction. Such an analysis, which the meteorologists in India successfully used in examining the performance of the prediction formulae, serves the additional purpose of convincing the applied worker of the validity of the technique. It is worth undertaking it specially in situations where the observations are taken over time or space or in accordance with some other factor.

Application in sampling surveys (Double sampling). An important use of multiple regression is in sample surveys where the object is the estimation of an average of a characteristic y in a population, by sampling for a number of related and easily ascertainable characteristics $x_1, x_2, \ldots$. When the regression of y on the related characteristics is unknown, a relatively small sample may be examined for all the characteristics to estimate the regression function. If $y = a+b_1x_1+b_2x_2+\ldots$ is the true or estimated (on the basis of a preliminary sample) regression equation and $t_1, t_2, \ldots$ are the estimated mean values of $x_1, x_2, \ldots$ based on a larger survey using any design, then the estimate of the mean of y is $a+b_1t_1+b_2t_2+\ldots$. Mahalanobis (1941) used this technique in forecasting the yield of cinchona bark by taking concomitant measurements on standing trees, such as their height, girth and thickness of the trunk at specified heights.

Estimation from fragmentary data. Rao and Shaw (1948) used the regression of cranial capacity on length, height and breadth of skulls to show that smaller skulls have a greater chance than bigger skulls of being recovered intact in excavations of old burial grounds for skeletal material. This suggests that the average cranial capacity of well preserved skulls underestimates to some extent the mean value for the skeletal population as a whole. Concomitant measurements such as length, height and breadth which could be taken even on broken skulls would provide, by the regression technique, a more representative estimate of the mean cranial capacity.

Selection problems (single stage, single and differential predictors). There is a group of problems such as industrial selection (recruitment for different jobs), admission of students to universities and institutions for higher education, vocational guidance, psychiatric screening of military recruits, selection of high yielding plants, etc., where the regression technique enables the choice of a given quota of individuals with the maximum expected 'payoff' which is the criterion variable expressed in money terms. A very general formulation of these problems is as follows.

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ represent the p.v.'s and $y_1, y_2, \ldots, y_s$ the payoffs corresponding to s criterion variables. The regression of y_i on $\boldsymbol{x}$ is denoted by $R_i(\boldsymbol{x})$. Let us find a region w in the space of $\boldsymbol{x}$ such that

$$\int_w dF(\boldsymbol{x}) = \alpha \text{ (given)} \qquad \ldots \quad (3.1)$$

$$\int_{w \times y_i} y_i \, dF(y_i, \boldsymbol{x}) = \int_w R_i(\boldsymbol{x}) \, dF(\boldsymbol{x}) \geqslant c_i \text{ (given)} \qquad \ldots \quad (3.2)$$

$$i = 2, \ldots, k \leqslant s$$

and

$$\int_w R_1(\boldsymbol{x}) \, dF(\boldsymbol{x}) \text{ is a maximum.} \qquad \ldots \quad (3.3)$$

By an application of the Neyman-Pearson lemma, we find that w is of the form

$$R_1(\boldsymbol{x}) \geqslant a_0 + a_2 R_2(\boldsymbol{x}) + \cdots + a_k R_k(\boldsymbol{x}) \qquad \ldots \quad (3.4)$$

in which case the equations (3.1, 3.2, 3.3) reduce to, say,

$$\left.\begin{array}{l} g_0(a_0, a_2, \ldots, a_k) = \alpha \\ g_i(a_0, a_2, \ldots, a_k) \geqslant c_i, \ i = 2, \ldots, k \\ g_1(a_0, a_2, \ldots, a_k) \text{ is a maximum.} \end{array}\right\} \qquad \ldots \quad (3.5)$$

The determination of $a_0, a_2, \ldots, a_k$ is then a programming problem generally nonlinear in nature. If, however, the inequalities in (3.2) are replaced by equalities then the coefficients $a_0, a_2, \ldots, a_k$ are determined from the equations

$$\left.\begin{array}{l} g_0(a_0, a_2, \ldots, a_k) = \alpha \\ g_i(a_0, a_2, \ldots, a_k) = c_i, \ i = 2, \ldots, k. \end{array}\right\} \qquad \ldots \quad (3.6)$$

We thus have a formal solution to the problem of selecting on the basis of p.v.'s a given quota of individuals in the long run, so as to maximise the expected payoff with respect to a certain characteristic, subject to the condition that the expected payoffs for certain other characteristics are not less than preassigned values. If there are no restrictions such as (3.2) the selection rule reduces to (Cochran, 1951; Roy and Mitra, 1954)

$$R_1(\boldsymbol{x}) \geqslant a_0.$$

An alternative problem is that of selecting individuals for different criteria, in given proportions, so as to maximise the total payoff. Let $\alpha_1, \alpha_2, \ldots$ be the proportions for various criteria. The problem is to find regions $w_1, w_2, \ldots$ so as to maximise

$$\int\limits_{w_1} R_1(\boldsymbol{x})\, dF(\boldsymbol{x}) + \int\limits_{w_2} R_2(\boldsymbol{x})\, dF(\boldsymbol{x}) + \ldots$$

subject to the conditions

$$\int\limits_{w_1} dF(\boldsymbol{x}) = \alpha_1, \int\limits_{w_2} dF(\boldsymbol{x}) = \alpha_2, \ldots \qquad \ldots \quad (3.7)$$

Using a generalisation of Neyman-Pearson lemma (Rao, 1948a, 1951a) the regions are found to be

$$w_i : R_i(x) + \lambda_i \geqslant R_j(x) + \lambda_j, \; j = 1, 2, \ldots \qquad \ldots \quad (3.8)$$

where $\lambda_1, \lambda_2, \ldots$ are determined to satisfy the equations (3.7).

The solutions (3.4) and (3.8) are valid only in the long run. In practice one has to select from a given number N of applicants, on whom values of the p.v.'s can be ascertained. Denoting by R_{ij} the expected value of the i-th c.v. given the p.v.'s of the j-th individual we consider the payoff matrix :

Individual no.	Expected payoff for the criterion variables
1	$R_{11} \;\; R_{21} \; \ldots \; R_{s1}$
.	
.	$. \quad . \quad \ldots$
N	$R_{1N} \;\; R_{2N} \; \ldots \; R_{sN}$

If n individuals are to be chosen to maximise the sum of $R_{1\lambda}$ subject to the condition that the sums of $R_{2\lambda}, \ldots, R_{k\lambda}$ are not less than preassigned values, the selection rule is to choose the i-th individual if and only if

$$R_{1i} \geqslant a_0 + a_2 R_{2i} + \ldots + a_k R_{ki} \qquad \ldots \quad (3.9)$$

where $a_0, a_2, \ldots, a_k$ are chosen to satisfy the stated restrictions on the sums, which is a problem of linear programming. It may be possible that no solution exists, in which case the conditions have to be altered so as to make them consistent (yield a solution for $a_0, a_2, \ldots, a_k$), or the constants as determined by equations (3.4) in the 'long run case' may be used, choosing $\alpha = n/N$.

If the problem is one of differential prediction (Brogden, 1951) i.e. of dividing N applicants into s groups of $n_1, n_2, \ldots, n_s$ individuals for the s criteria, then the i-th individual is assigned to the j-th criterion if and only if

$$R_{ji} + \lambda_j \geqslant R_{qi} + \lambda_q, \; q = 1, \ldots, s \qquad \ldots \quad (3.10)$$

where λ_j are chosen such that the procedure (3.10) leads to given numbers of individuals for the various criteria. This is again a particular case of linear programming; however a simple geometrical solution exists in the case of three groups (Rao, 1952, p.328).

One important problem to be considered is the estimation and test of significance of selective efficiency[1] of the p.v.'s for a criterion variable y_1 when some restrictions are placed on the expected values of other c.v.'s in the selected group. This problem has been recently solved by the author and the results are being published.

Sequential (multistage) selection procedures. All the p.v.'s may not be available in a single stage, at a given point of time, to decide on the selections. It may be convenient then to adopt a sequential method. At each stage when a fresh set of p.v.'s is available we may, in any individual case, select, reject or postpone decision till further observations are obtained. At the last stage when no more p.v.'s are intended to be observed there are only two possible decisions, viz., selection or rejection. Such a sequential procedure of selection results in considerable economy as the cost of obtaining information on the p.v.'s at subsequent stages is confined only to those provisionally selected at the previous stage.

To specify a sequential procedure we may use certain quotas of selection and rejection at each stage. We shall consider the simple, but important case of two-stage selection, where, in the first stage a given proportion of individuals is provisionally accepted and from among those provisionally accepted another given proportion is selected at the second stage, by using further information.

Let us denote the first stage p.v.'s by $\boldsymbol{x} = (x_1, x_2, \ldots)$ and the second stage, by $\boldsymbol{u} = (u_1, u_2, \ldots)$. The proportion to be selected in the first stage is β and both stages together is α. If w_1 denotes a region in the $\boldsymbol{x}$ space for first stage selection, w_2 a region in the product space of w_1 and $\boldsymbol{u}$ for the second stage selection, and $R(\boldsymbol{x}, \boldsymbol{u})$ is the regression of the criterion variable on $\boldsymbol{x}$ and $\boldsymbol{u}$, then the problem is to maximise

$$\int_{w_2} R(\boldsymbol{x}, \boldsymbol{u})\, dF(\boldsymbol{x}, \boldsymbol{u}) \qquad \ldots \quad (3.11)$$

subject to the conditions

$$\int_{w_1} dF(\boldsymbol{x}) = \beta \quad \text{and} \quad \int_{w_2} dF(\boldsymbol{x}, \boldsymbol{u}) = \alpha$$

It is easy to see that w_2 is defined by $R(\boldsymbol{x}, \boldsymbol{u}) \geqslant \lambda$ where $\boldsymbol{x} \epsilon w_1$, so that if w_1 is given, w_2 is completely specified. For given λ let

$$\int_{w_2} R(\boldsymbol{x}, \boldsymbol{u})\, dF(\boldsymbol{x}, \boldsymbol{u}) = \int_{w_1} f_0(\boldsymbol{x}, \lambda)\, dF(\boldsymbol{x}) \qquad \ldots \quad (3.12)$$

$$\int_{w_2} dF(\boldsymbol{x}, \boldsymbol{u}) = \int_{w_1} f_1(x, \lambda)\, dF(x) = \alpha, \quad \int_{w_1} dF(x) = \beta \qquad \ldots \quad (3.13)$$

Then the best choices of w_1, w_2 for given λ are

$$\left.\begin{aligned} w_1 &: f_0(\boldsymbol{x}, \lambda) \geqslant a_1 f_1(\boldsymbol{x}, \lambda) + a_0 \\ w_2 &: R(\boldsymbol{x}, \boldsymbol{u}) \geqslant \lambda \end{aligned}\right\} \qquad \ldots \quad (3.14)$$

---

[1] Selective efficiency of a procedure is defined as the increase in the mean value of the selected group over that in the case of random selection.

where a_1 and a_0 are determined to satisfy the equalities (3.13). We can then choose λ to maximise (3.12). The actual solution is thus extremely complicated although some simplification is possible when the variables have a multivariate normal distribution. Cochran (1951) suggested the regions

$$w_1 : R(\boldsymbol{x}) \geqslant \mu, \quad w_2 : R(\boldsymbol{x}, \boldsymbol{u}) \geqslant \lambda$$

which, no doubt, lead to a simple procedure of selection, but differ considerably from the optimum solution (3.14) as may be expected.

As in earlier problems, if to begin with there are N individuals providing information on the first set of p.v.'s, n_1 are to be provisionally accepted at the first stage and n_{12} are to be ultimately selected, the solution is the same as in (3.14), except that the constants are determined such that the procedure selects only the specified numbers of individuals, n_1 and n_{12}, at the two stages.

Formal solutions may be obtained for other sequential procedures, but the computations needed would be extremely heavy.

Prediction in a stationary series. In the examples considered earlier a qualitative distinction could be made between the p.v.'s and c.v. in that they represented the measurements of different characteristics. But this need not be so as when we would like to predict the price of rice next month on the basis of the prices in the previous months, or the height of the next individual to be drawn from a group given previously observed heights of a number of individuals drawn at random from the same group, or the number of defectives in a sample of a given size from the next lot of articles supplied for inspection on the basis of knowledge available from the previous lots. In all such cases the p.v.'s are what have been observed in the past and the c.v. a future observation on the same characteristic. The problem is, in general, that of estimating the distribution function of future observations, or simply some characteristic of the distribution such as the mean, or approximations of given functional types to the mean etc. Satisfactory solutions are available by the method of Fiducial Probability (Fisher, 1956), in situations where the conditions for its application are satisfied as in the problem of predicting the height of the next individual to be drawn, or the number of defectives in a sample from a given lot. In other cases large sample methods are available.

An important application is prediction in a time series which constantly arises in economic, scientific and technological investigations such as the prediction of future price, meteorological forecasts, random noise in communication theory, turbulence, storm-generated ocean waves etc. There is considerable literature on the subject originating with the works of Kolmogoroff (1941) and Wold (1938). The problem is one of fitting a linear regression function of x_h, the variable after time h, on a large (theoretically infinite) number of previous variables $x_{-1}, x_{-2} \ldots$. Traditional methods are not applicable as the inversion of an infinitely large covariance matrix is involved. It is, however, possible to obtain a prediction formula involving an infinite number of previous variables, which could be used as an approximation when a large series is available, under some conditions on the underlying stochastic process. One main condition is stationarity or weak stationarity which specifies that the covariance between x_r and x_s is a function of the difference $(r-s)$ only. For other conditions and detailed computations reference may be made to the book on stochastic processes by Doob (1952). Recently Wiener and Masani (1958) generalised the results to multivariate stochastic processes. In particular they considered the simultaneous prediction of a variable at a number of future time points.

4. DISCRIMINANT FUNCTION AND DISCRIMINATORY TOPOLOGY (C.V.'S ON NOMINAL AND P.V.'S ON INTERVAL SCALE)

The theory of the discriminant function arose out of a procedure suggested by R. A. Fisher for sorting a collection of jaw bones according to sex of the owner. The problem is one of predicting sex (male or female) of the owner in each case from a number of measurements on the jaw bone. In such a situation, the criterion variable viz., sex, has only a nominal scale while the predictor variables may have an interval scale.

The formal mathematical solution was supplied by Welch (1939) in the case of classification of an individual into one of two alternative populations. Let $\boldsymbol{x} = (x_1, x_2 \ldots)$ denote the p.v.'s and $f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$ the probability densities for $\boldsymbol{x}$ in the two populations, and π_1, π_2 the relative frequencies of individuals of the two populations. By applying Bayes' theorem on conditional probabilities we find that the *a posteriori* probabilities for an individual with measurements $\boldsymbol{x}$ to belong to the first and second populations are proportional to $\pi_1 f_1(\boldsymbol{x})$ and $\pi_2 f_2(\boldsymbol{x})$ respectively. To minimise the errors of classification the Bayes solution is to assign the individual to the first population if

$$\pi_1 f_1(\boldsymbol{x}) > \pi_2 f_2(\boldsymbol{x}), \quad \ldots \quad (4.1)$$

and to the second population if

$$\pi_2 f_2(\boldsymbol{x}) > \pi_1 f_1(\boldsymbol{x}), \quad \ldots \quad (4.2)$$

and any random assignment may be followed in the case of equality.

We can consider, if necessary, the losses involved in wrong decisions and minimise the expected loss. If r_{ij} stands for the monetary return in classifying an individual of the i-th population to the j-th population, then the regression of the payoff (or monetary gain) when an individual with measurements $\boldsymbol{x}$ is classified as a member of the first population is proportional to (using *a posteriori* probabilities)

$$r_{11}\pi_1 f_1(\boldsymbol{x}) + r_{21}\pi_2 f_2(\boldsymbol{x}) \quad \ldots \quad (4.3)$$

and for the second it is proportional to

$$r_{12}\pi_1 f_1(\boldsymbol{x}) + r_{22}\pi_2 f_2(\boldsymbol{x}). \quad \ldots \quad (4.4)$$

Following the concept of differential predictors of Section 3 (multiple regression), where an individual is selected for a job where his expected payoff is a maximum, the individual is assigned to the first population if (4.3) $>$ (4.4) and otherwise to the second.[1]

The extension of this solution to classification of an individual as a member of one of several alternative populations is again a simple application of Bayes' theorem and regression function; explicit solutions are given in (Rao, 1948a, 1950, 1951a, 1951b, 1952).

Although the formal solutions are simple there are a number of practical difficulties in their application. We shall consider a few examples and examine how the problems are solved in practice.

Differential diagnosis of neurotic states. Diagnosis of different types of neurotic diseases such as Anxiety state, Hysteria, Psychopathy, Obsession, Personality change and absence of any abnormality is extremely complicated in practice. Any mechanisation at least in distinguishing between states not close to one another in mental deterioration will be useful. A number of psychological tests administered to patients, previously classified into these states, showed differences in the response patterns and threw out the possibility of using an individual's response pattern in differential diagnosis. Using the estimates of their mean values, and variances and covariances the probability density of the test scores, say, $x_1, x_2, \ldots,$ for any neurotic group or a normal group can be computed. The relative frequencies of these patients and normal individuals coming for diagnosis will also be available from past experience. We thus have all the pre-requisites for an application of the discriminant function. Rao and Slater (1949) showed that a routine application of the method involves the computation of a linear discriminant score on an individual for each of the possible categories and assigning him to the category with the highest discriminant score.

There are other clinical applications such as the discrimination between two types of Jaundice, one of which needs a medical treatment and the other surgery. It has been found that two biochemical tests carried out with very little expense provide a discriminant function for classifying the patients into one of the two groups. The discriminant function can thus provide an objective criterion for coming to decisions. It is not suggested that other observations of diagnostic value should not be taken. It is only claimed that in routine work where the resources enable only a limited number of tests to be carried out on an individual there exists an optimum procedure which minimises the errors of misclassification or the overall risk if magnitudes of risk due to different types of wrong classification can be specified.

Differential diagnosis, Doubtful region. The procedure outlined above no doubt reduces the probability of misclassification to the minimum possible, but if this is large we do not have sufficient assurance in individual cases. However, whatever may be the measurements available, it should be possible in some cases to come to a decision, subject to a small preassigned risk, whereas in other cases either no decision or only a provisional decision could be made. The latter comprises the doubtful cases which need further examination. The rules for such decisions can be formulated on the basis of the likelihood ratio as in the earlier problem. Details of such sequential procedures are given in (Rao, 1952).

---

[1] The solution (4.1, 4.2) refers to the special case where the payoff for correct classification is 1 and for wrong classification zero. In the general case the problem is essentially an application of multiple regression and differential prediction.

Classification of Fossils. In problems of discrimination such as sexing of skeletal remains or the diagnosis of type of Jaundice, the possibilities of identification are generally limited to two. However, where the external evidence is slight or equivocal, the assignment of an individual to one of two groups may be subject to another type of error, viz., the wrong assumption that he belongs to one of the two groups when, in fact, he comes from a third unknown group. In the absence of any definite knowledge, we may have to provide for the possibility of not accepting the individual for either of the given groups.

Consider, for example, *Australopithecus africanus*, a fossil (consisting of the face, part of the base of the skull, the jaws and teeth) discovered in Africa and described in a letter to Nature by Dart (1925). There is considerable controversy (Dart, 1934; Ashton and Zuckerman, 1950; Bronowski and Long, 1951; Yates and Healy, 1951) whether *A. africanus* is more 'hominid' than 'anthropoid'. Comparisons are made, for this purpose, between the measurements of fossil teeth with those of men and the anthropoid apes. Bronowski and Long (1951) computed a linear discriminant function between chimpanzees and men on the basis of four measurements on teeth and found that *A. africanus* is not significantly different from men but significantly different from chimpanzees with respect to the discriminant function. They concluded that *A. africanus* is more hominid. The procedure followed is efficient only when it is known that the fossil belongs either to one of the two groups considered or corresponds to some species with mean values lying on a linear path (evolutionary?) connecting the mean values of men and chimpanzees. We may, therefore, approach the problem by first testing the latter hypothesis, which when not rejected by the data may justify the use of the discriminant function for coming to a decision. On the other hand, if this hypothesis is rejected, or if with respect to the discriminant function the measurement on the fossil is significantly different from the means of the two groups we have evidence that the discovered fossil possibly belongs to a third group.

If μ^1, μ^2 are the vectors of mean values for the given groups and Λ is the common dispersion matrix and $\boldsymbol{x}$ is the vector of measurements on the fossil, then the χ^2 on p degrees of freedom for testing the deviation of $\boldsymbol{x}$ from one of the groups with mean (say) μ^1 is

$$(\boldsymbol{x}-\mu^1)' \; \Lambda^{-1}(\boldsymbol{x}-\mu^1), \qquad \ldots \quad (4.5)$$

and this can be decomposed into two components

$$\frac{\{(\boldsymbol{x}-\mu^1)'\Lambda^{-1}(\mu^1-\mu^2)\}^2}{(\mu^1-\mu^2)'\Lambda^{-1}(\mu^1-\mu^2)}+\chi_2^2 \qquad \ldots \quad (4.6)$$

the first with 1 d.f. corresponding to the deviation with respect to the discriminant function and the other with $(p-1)$ d.f., to directions independent of the discriminant function. The second component χ_2^2 provides a valid test of the hypothesis that the discriminant function supplies all the information in coming to a decision. Using the computations given by Yates and Healy (1951) for the length and the breadth of tooth and adopting a reasonable value of $+0.3$ for the correlation coefficient between length and breadth,[1] the value of χ_2^2 on $(p-1) = 1$ d.f. is of the order of 5, which is significant at the 5 percent level. This shows that (i) *A. africanus* perhaps belongs to a third species, a conclusion which the multivariate test leads to more definitely than the univariate approach[2] of Yates and Healy (1951), and (ii) the use of the discriminant function alone, by which Bronowski and Long (1951) concluded that *A. africanus* is not significantly different from men, can be misleading.

It may be noted that the χ^2 distribution we used in the problem of *A. africanus* is valid on the assumption that the mean values and the dispersion matrix of the two groups are well determined. The theory can, however, be applied even in small samples; the appropriate distributions in this case depend on the distribution of roots of certain determinantal equations. For the theoretical developments reference may be made to Fisher (1939), Hsu (1939) and Roy (1939). Associated problems and applications are contained in papers by Bartlett (1951), Rao and Slater (1949), Williams (1952) and others.

Vocational guidance. In which of the vocations $A, B, C, \ldots$, will an individual with given scores in a battery of tests have the maximum chance of being successful? Such a problem is sometimes considered as one of discrimination; i.e., of classifying an individual as a member of one of the vocation groups[3] A, B, C, ..., using the methods discussed above. But the logical situation in vocational guidance seems to be slightly different. Our attempt should be to compute for any individual with given test scores the chances of his being successful in each of a number of vocations. He could then choose a vocation in which his chance of success is a maximum.

Let $p_g(\boldsymbol{x})$ be the probability density of the test score $\boldsymbol{x}$ in the general population; $p_A(\boldsymbol{x})$ for those successful in vocation A; $p_B(\boldsymbol{x})$ for those successful in B and so on. If $s_i(x)$ is the probability of success in vocation i for an individual with test scores $\boldsymbol{x}$, then

$$p_i(\boldsymbol{x}) \propto p_g(\boldsymbol{x})s_i(\boldsymbol{x}), \quad i = A, B, \ldots \qquad \ldots \quad (4.7)$$

Hence $s_i(\boldsymbol{x}) \propto p_i(\boldsymbol{x})/p_g(\boldsymbol{x})$, the likelihood ratio of the i-th vocation to the general population, so that if a number of individuals have to be ordered by the probability of success in vocation i we can use $s_i(\boldsymbol{x})$ or $p_i(\boldsymbol{x})$ itself as an index.

---

[1] I have not been able to find out the value of the correlation coefficient used by Yates and Healy (1951). The conclusion is not altered even if the value is much smaller than $+0.3$.

[2] Yates and Healy found that the fossil is significantly different from chimpanzees in each of the two measurements and from men only in one of the measurements.

[3] A vocation group is the group of individuals who are successful in a particular profession.

To compare $s_i(\boldsymbol{x})$ over the vocations we need to know the constant of proportionality of $s_i(\boldsymbol{x})$ to the ratio $p_i(\boldsymbol{x})/p_g(\boldsymbol{x})$. If π_i is the ratio of those successful to the total number of persons in vocation i then

$$s_i(\boldsymbol{x}) = \pi_i p_i(\boldsymbol{x})/p_g(\boldsymbol{x}). \qquad \dots \quad (4.8)$$

If π_i are known, then the chances of success over the vocations are proportional to $\pi_i p_i(\boldsymbol{x})$, which resemble discriminant scores in the discriminatory analysis with π_i representing *a priori* probabilities of the various groups. The vocation with the highest chance of success is that for which $\pi_i p_i(\boldsymbol{x})$ is a maximum.

The equation (4.8) provides a satisfactory solution to the problem of vocational guidance if at least approximate values of π_i are ascertained. This approach differs in the concepts involved from those found in the literature on vocational guidance.

Mahalanobis' D^2 and discriminatory topology. While considering the problem of classification of an individual into one of two given groups a question was raised regarding the extent to which errors of misclassification could be reduced by increasing the number of measurements. When the measurements follow a normal distribution it can be shown (Rao, 1948a) that the minimum probability of misclassification using p characters is a decreasing function of Mahalanobis' distance D_p^2, defined by

$$D_p^2 = \Sigma\ \Sigma\ \alpha^{ij}\delta_i\delta_j$$

where δ_i is the difference in the mean values of the i-th measurement for the two groups and (α^{ij}) is the matrix inverse to the dispersion matrix. The question posed can then be answered by determining the limit D_∞^2 of D_p^2 as $p \to \infty$. If $D_\infty^2 < \infty$, there is a non-zero[1] lower limit to probabilities of misclassification and therefore, not much gain results by indefinitely increasing the number of measurements.

It is, therefore, of great interest to examine models of multivariate distributions with infinitely many dimensions and having finite values of D_∞^2 between groups.

Let us imagine hypothetical variables $\eta_1, \eta_2, \dots$ all uncorrelated and with unit variance and which may be denumerably infinite in number. The variables may be thought of as the hypothetical effects of major genes or polygenic systems (Mather, 1949) in which case they may be called genotypic variables. An observed measurement x_i may be considered as consisting of two parts, one determined by linear combinations of the genotypic variables and another a random variable, representing the effect of the environment. Thus

$$x_i = \Sigma a_{ji}\eta_j + \epsilon_i$$

Theoretically the index i may belong to any set, denumerable or not, thus admitting variables defined even with respect to a continuous set of coordinates.

---

[1] Whatever may be the distribution of the variables, the probability of errors of classification $\to 0$ as $D_p^2 \to \infty$, but it cannot be said that the limit is non-zero when D_p^2 does not diverge except in the special case of normal distributions.

Let D_p^2 be the Mahalanobis' distance between two groups based on a subset p of measurement x_i. When the effect of the environment is the same for the two groups, it can be shown by extending the arguments used earlier by the author (Rao, 1954b) that

$$\underset{p\to\infty}{\text{limit}} \left\{ \begin{array}{c} \text{supremum } D_p^2 \\ \text{(over sets of } p) \end{array} \right\} \leqslant \Sigma\delta_i^2$$

where δ_i is the difference in mean values of the genetic variable η_i. We thus have a sufficient condition for convergence of D^2, viz., $\Sigma\delta_i^2$, the genetic distance between the groups be finite, which is true when the number of genetic variables in which the two groups differ is finite or when the differences in genetic variables, except for a finite number, are negligibly small so as to make the infinite sum of squares finite.

The convergence of the genetic distance opens out the possibility of studying relationships among given sets of groups based on genetic similarity or divergence by examining only a few measurements. With a suitably chosen small number of measurements it is possible to obtain a good idea of the underlying genetic distance with respect to which a classification of the groups is sought.

This line of work regarding classification of groups, originally introduced by Mahalanobis, has been fruitfully pursued by the workers at the Indian Statistical Institute and elsewhere and the methods developed have been successfully applied in a wide variety of situations (Blackith, 1957; Karve and Dandekar, 1951; Mahalanobis, 1928, 1930, 1936; Mahalanobis, Majumdar and Rao, 1949; Majumdar and Rao, 1958; Mukherjee, Rao and Trevor, 1955; Nair, 1952).

General remarks. It has been generally assumed, in the derivation of the methods for discrimination, that the population parameters are known or estimated from a fairly large sample. Some appropriate procedures when parameters are obtained from small samples have been given by the author (Rao, 1954a). A deeper probe into this problem is necessary. In a recent paper John (1959) has examined the errors of misclassification for certain procedures taking into account the errors in estimation of the parameters.

5. PREDICTION OF THE UNOBSERVABLE (HYPOTHETICAL VARIABLES)

We consider here problems of regression when the c.v. has an interval scale but is not observable. For instance, when a bull or a plant is chosen for further propagation, the c.v. is the expected performance of the progeny, which indeed is not directly measurable. Or we may like to grade a number of individuals with respect to a certain aptitude for which no direct measurement is available but which is reflected in the performance of an individual in a battery of tests. In such cases we postulate the existence of some variables, though unobservable, and predict their values through a study of their dependence on the observable variables.

Discriminant function for genetic selection. When quantitative characters are involved, it is considerably difficult to select genetically desirable types in breeding work because heritable differences are to some extent masked by non-heritable or environmental variations. The problem then arises as to what is the best indicator of the genotypic value of any individual line. Let $(\psi_1, \psi_2, \ldots)$ be the genotypic values of certain characteristics (like the annual number of eggs laid, the size of the egg, etc. in the case of poultry) and $(x_1, x_2, \ldots)$ the observed values of these characteristics on an individual of a progeny; our object is to predict $(\psi_1, \psi_2, \ldots)$ for a line from the observations on a number k of individuals of the progeny. Assuming linear regression it can be shown that the prediction formula for ψ_i is of the form

$$\psi_i = \alpha_i + \beta_{1i}\bar{x}_1 + \ldots + \beta_{pi}\bar{x}_p = R_i(\bar{x}) \qquad \ldots \quad (5.1)$$

where $\bar{x}_j$ denotes average over k individuals of the progeny for the j-th character. The computations involved in the estimation of α_i and the β coefficients are discussed in a paper by the author (1953), where it was pointed out that the estimates depend on the sample size k available for each line. To predict the genotypic values of different lines for purposes of comparison and selection it may be, therefore, necessary to use different sets of regression coefficients when the individuals observed for each line are not the same. This has not been explicitly mentioned in the literature including my earlier exposition of the subject (Rao, 1952). For instance, Panse (1946) and others used a single set of coefficients even in situations where the numbers of samples for the various progenies are different.

There are also a few points on the computational side which are worth stressing. Generally, what is needed is the prediction of a linear compound of $\psi_1, \psi_2, \ldots$ using weights according to their commercial value. If individual regressions $R_i(\bar{x})$, (equation 5.1) are computed in advance, then it is easy to see that the regression equation for the linear compound $(a_1\psi_1 + a_2\psi_2 + \ldots)$ is simply $(a_1R_1(\bar{x}) + a_2R_2(\bar{x}) + \ldots)$. Such a procedure, instead of directly estimating the regression of $(a_1\psi_1 + a_2\psi_2 + \ldots)$ on $\bar{x}_1, \bar{x}_2, \ldots$, has some flexibility in that it allows for an easy adjustment of the prediction function if it is desired to alter the commercial values, or to compute functions for different sets of commercial values.

Another advantage in obtaining $R_i(\bar{x})$ to begin with, which can be done without heavy computations, is in the adjustment needed in the selection index under some restrictions such as, that selection should not alter the expected values of certain other genotypic characteristics. Let us represent by $(a, \psi) = [a_1\psi_1 + a_2\psi_2 + \ldots]$ the linear compound we want to maximise, and by (b, ψ), (c, ψ), ... the other linear compounds in which no change is desired. The solution to such a problem of selection under restrictions is derived in Section 3 by an application of the Neyman-Pearson lemma. It was shown that if $(a, R) = [a_1R_1(\bar{x}) + a_2R_2(\bar{x}) + \ldots]$ is the selection index for (a, ψ), then the selection index under given restrictions is

$$(a, R) - \lambda_1(b, R) - \lambda_2(c, R) + \ldots \qquad \ldots \quad (5.2)$$

where $\lambda_1, \lambda_2, \ldots$ are chosen such that the covariances of (5.2) with (b, R), (c, R), $\ldots$ are all zero. The equations for $\lambda_1, \lambda_2, \ldots$ are

$$\begin{aligned} \text{cov}\,[(a, R), (b, R)] &= \lambda_1 \,\text{cov}\,[(b, R), (b, R)] + \lambda_2 \,\text{cov}\,[(b, R), (c, R)] + \ldots \\ \text{cov}\,[(a, R), (c, R)] &= \lambda_1 \,\text{cov}\,[(c, R), (b, R)] + \lambda_2 \,\text{cov}\,[c, R), (c, R)] + \ldots \\ \ldots \quad & \quad \ldots \quad \quad \ldots \end{aligned}$$

The procedure outlined above appears to be somewhat simpler than the application of formulae given by Kempthorne and Nordskog (1959) in the problem of restricted selection indices.

General remarks. A straight selection function $(a_1\bar{x}_1 + a_2\bar{x}_2 + \ldots)$ for the linear compound $(a_1\psi_1 + a_2\psi_2 + \ldots)$ is sometimes recommended when the extra computations involved cannot be undertaken or when the estimates of genetic and environmental variances and covariances are not very reliable or when there is very little extra gain by using the best possible index. An important statistical tool needed in this connection is a test of the hypothesis that the straight selection function is the best index. An exact solution is supplied when the genetic covariance matrix has rank unity, and only an approximate solution in other cases (Rao, 1953). Further research is needed in the solution of this problem.

Another line of investigation which needs further research is the theory of familial correlations obtained by generalising the univariate intra-class correlation to multiple measurements (Rao, 1945). The nature of the genotypic variables $(\psi_1, \psi_2, \ldots)$ and its association with the corresponding observable variables on the progeny $(x_1, x_2, \ldots)$ can be studied with the help of familial correlations.

Factor analysis. The statistical procedures discussed in the previous sections are concerned with the best utilization of any given set of measurements such as the scores in a battery of tests in predicting a criterion variable. Given several alternative batteries of tests it is, however, possible by appropriate statistical analysis to choose the set which is most efficient for prediction. But the construction of alternative batteries of tests or the procedure for improving a battery of tests is itself a problem which has to be faced in practice. It is claimed that factor analysis would be of some use in this direction. Whether factor analysis by itself would be a proper guide in such cases is, however, a debatable issue.

Further, it is said that for predicting a criterion variable, factor scores extracted from test scores may be used. This is clearly an inefficient procedure, for the test scores themselves would be better predictors.

In order to understand how exactly factor analysis may be useful let us examine the problem as formulated by the psychologists. If $x_1, \ldots, x_p$ represent the test scores on an individual then the model considered is

$$x_i = a_{i1}f_1 + \ldots + a_{ik}f_k + s_i + \epsilon_i \qquad \ldots \quad (5.3)$$
$$i = 1, \ldots, p$$

where $f_1, \ldots, f_k, s_i, \epsilon_i$ $(i = 1, \ldots, p)$ are all uncorrelated hypothetical variables, s_i representing a variable specific to x_i, ϵ_i a random error and $f_1, \ldots, f_k$, each of which has non-zero coefficients for at least two of the variables x_i, representing common factors. The coefficients a_{ij} are unknown. It is easy to see that the correlations among x_i are solely due to the common factors which, as they appear in model (5.3), are neither unique nor operationally defined. We can make any orthogonal transformation of the factors and get a structure similar to that of (5.3). The problem differs in this respect from that of discriminant function for genetic selection where the hypothetical variables are defined as expectations of certain observable random variables.

Making use of the theory of canonical correlations developed by Hotelling (1936), a procedure called canonical factor analysis (CFA) was given by author (Rao, 1955a) for estimating a basis of the *whole space* of the common factors with the minimum rank necessary to explain the intercorrelations apart from sampling errors. This enables the prediction of any element of the common factor space provided it is specified in terms of the basis of the space (also Lawley, 1940, 1942).

If factor analysis is to be considered as a method of predicting a hypothetical common factor, it is applicable only in situations where (i) an external criterion is not defined, (ii) the criterion in which we are interested is an element of the common factor space and (iii) this element is somehow identified and suitably specified in terms of the basis characterising the common factor space. There is nothing to guide us in the statistical technique of factor analysis regarding (ii) and (iii); the considerations have to be extra-statistical.

6. MULTIVARIATE TESTS OF SIGNIFICANCE

The construction of a prediction formula in any given situation involves prior statistical analysis of past data involving tests of hypotheses concerning the relationship between the p.v.'s and c.v.'s and estimation of unknown parameters. Typical hypotheses to be examined are : (i) Whether a given set of p.v.'s provide information on the c.v.; (ii) Whether addition of other variables to a given set of p.v.'s improves the prediction of the c.v.'s and; (iii) Is a specified relationship between the p.v.'s and c.v. true ?

The theory of least squares for testing linear hypotheses can be used in the multiple regression situation. Most of the multivariate tests developed have special reference to problems considered under the category of discriminant functions. Research in this direction originated with Wishart's (1928) distribution of the estimated variances and covariances and Hotelling's T^2 test (1931) and canonical correlations (1936), some of which were generalised by Wilks (1932, 1946). Fisher (1939) broke new ground with the introduction of test statistics for examining the dimensionality of the configuration of mean vectors of a number of populations. Similar test criteria were developed by Roy (1939) for testing equality of variance-covariance matrices of two populations. Two important papers concerning non-null distributions are by

Fisher (1928) on the multiple correlation coefficient and by Bose and Roy (1938) on Mahalanobis' D^2. Approximations to complicated distributions of some multivariate test criteria are given by Bartlett (1939a), Box (1949), Rao (1948b, 1951c), J. Roy (1951) and others.

The problem of testing improvement in prediction by the addition of more predictor variables was examined by Rao (1946, 1948b, 1949) generalising a test proposed by Fisher (1940) for a hypothetical discriminant function (also Bartlett, 1939b, 1947). For further contributions useful in the general study of problems considered in the present paper reference may be made to books by Anderson (1958), Kendall (1954) and Rao (1952). Applications of multivariate tests are also contained in other papers by the author (Rao, 1954c, 1955b, 1956, 1958).

Multivariate tests are sometimes criticised on the ground that they are designed to test only overall differences, which are difficult to interpret, whereas the main interest often would be in examining individual differences in a few specified cases. It is argued that since univariate methods are admirably suited to the latter purpose, there is no need to indulge in the luxury of more complicated techniques, or to carry out a test which can detect differences of any sort and provide a logical basis for interpreting the individual differences.

It is, however, a wrong attitude to take that data should be examined only for the specific purpose for which they are collected, or that data collected should be just sufficient to answer problems of immediate interest. This is surely detrimental to scientific progress. Only an elaborate analysis and more comprehensive data can insure the research worker against possible pitfalls in the interpretation of statistical tests and, what is more important provide clues for further investigation. The history of science tells us that several new discoveries made on the basis of evidence supplied by observed data were not envisaged by the investigator at the time of data collection.

7. A FEW OTHER PROBLEMS FOR RESEARCH

Nominal c.v. and nominal p.v.'s. Consider the problem of discrimination of three types of beetles, E_{12} Stigma, E_{28} Punctata and E Implicata on the basis of absence or presence of spots in each of 7 well defined positions on the thorax and/or 14 positions on the wing. If the presence of a spot is scored as '1' and absence as '0' arbitrarily, the scores of an individual can be represented as a vector $(a_1, a_2, \ldots)$ where each a can take values 0 or 1 with the order representing the positions. If the theoretical frequencies of the 2^7 possible vectors $(a_1, a_2, \ldots)$, in the case of 7 positions, are known for each of the species and if the relative abundance of members of these species is specified, then the Bayes solution can be used for discrimination. A major difficulty in this problem is the estimation of the frequencies of the $2^7 = 128$ combinations for each of the species from a preliminary sample. Since the number of unknowns is very large, 128 in the present case, no reliable estimates can be found without making any assumptions. Bahadur (1956) and Lazarsfeld (1956) have suggested different types of approximations to the theoretical frequencies involving the estimation of fewer parameters. To what extent these approximations are adequate in practical problems has yet to be examined.

Ordinal c.v. Suppose it was possible to obtain the scores in a battery of tests together with a ranking in performance on a particular job of a number of individuals. How can this information be utilised to predict the ranks for a fresh batch of individuals knowing only the test scores? If the underlying c.v. and the predictor variables can be considered to have a multivariate normal distribution, then it may be possible to construct the regression function apart from a constant by first estimating the necesssary unknown parameters from the given data. The expression for probability of the observations is extremely complicated and therefore the method of maximum likelihood may not be practicable. Some simpler methods may be explored. Another procedure is to quantify the ranks by substituting for the observed i-th percentile rank the corresponding theoretical normal deviate. With such quantities for the c.v. the multiple regression method may be applied. No reasonable solution seems to be available when nothing is known about the distribution of the underlying c.v.

A closely related problem is the estimation of the regression function when only the ranks of the predictor variables have been observed and the c.v. has been measured on an interval scale. Assuming a multivariate normal distribution for all the underlying variables it may be possible to obtain reasonable estimates of the regression parameters.

Another problem of interest is the estimation of structural relationships from observations subject to errors with specified or unspecified distributions. A closely related problem arises in the study of organic correlation and allometry, where a straight line is fitted to show the relationship between two or more variates. Contributions by Teissier (1948) and Kermack and Haldane (1950) have led to a proper understanding of the problem and the difficulties involved in statistical estimation and tests. Some applications to paleontological problems have also been suggested by Kermack and Haldane (1950).

Often observations are taken on individuals from mixed populations. How can they be used to recognise that the sampled population is possibly a mixture of several homogeneous groups and to estimate the parameters of each of the groups as well as the proportion of each group is to the mixture? Karl Pearson solved the problem of estimation when the mixture is of two normal distributions with different means and variances. Poti (1948) and Rao (1948a) worked the procedure for equal variances, in which case the solution depends on a root of a simple cubic instead of a nonic equation which arises in Karl Pearson's problem. There is need to develop statistical methods for estimation and tests of significance for mixtures of three or more populations and also of non-normal populations. The need is somewhat acute considering the fact that in most field investigations data collected refer to a heterogeneous population.

I am grateful to Professor J. B. S. Haldane for pointing out some problems which are analogous to those under discussion and which cannot be directly solved by methods discussed in this paper. One problem is the estimation of the frequencies of different genotypes (for example, 4 classes of genotypes A A B B C C, A a B B C C, A a B b C C, A a B b C c) among the parents knowing the numbers of dominant (i.e. having at least one dominant gene for each factor) and recessive children in each family. Another type of problem which is more difficult arises when the parents fall into even as few as two classes but the frequencies of progeny in each must be estimated from the data, a typical example of which is the data given by Bridges and others on primary and secondary non-disjunction.

In an attempt to describe to you some statistical procedures under the title, 'An indispensable aid in applied research,' I hope I have not disappointed those who are interested in applications, by over-emphasizing the imperfections of existing statistical tools or by reference to unsolved problems. I would perhaps be failing in my duty as president of this section if I did not indulge in self-criticism as well.

References

Anderson, T. W. (1958) : *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, New York.

Ashton, E. H. and Zuckerman, S. (1950) : Some quantitative dental characters of fossil anthropoids. *Philos. Trans. Roy. Soc.* (Series B), **234**, 485.

Bahadur, R. R. (1956) : A representation of the joint distribution of responses to n dichotomous items. School of Aviation Medicine. USAF report 59–42, 1959.

Bartlett, M. S. (1939a) : A note on tests of significance in multivariate analysis. *Proc. Camb. Phil. Soc.*, **35**, 180.

——— (1939b) : The standard errors of discriminant function coefficients. *J. Roy. Stat. Soc.* (Suppt.), **6**, 169.

——— (1947) : Multivariate analysis. *J. Roy. Stat. Soc.*, Series B, **9**, 176.

——— (1951) : The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Ann. Eugen.*, **16**, 190.

Blackith, R. E. (1957) : Polymorphism in some Australian locusts and grasshoppers. *Biometrics*, **13**, 183.

Brogden, H. E. (1951) : Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educ. Psychol. Measmt.*, **11**, 173.

Bronowski and Long (1951) : Statistical methods in anthropology. *Nature*, **168**, 794.

Bose, R. C. and Roy, S. N. (1938) : The exact distribution of the studentised D^2-statistic. *Sankhyā*, **4**, 535.

Box, G. E. P. (1949) : A general distribution theory for a class of likelihood ratio criteria. *Biom.*, **36**, 317.

Cochran, W. G. (1951) : Improvement by means of selection. *Second Berkeley symposium on mathematical statistics and probability.* University of California Press, 449.

Cronbach, L. J. and Gleser, C. G. (1957) : *Psychological Tests and Personnel Decisions.* University of Illinois Press, U.S.A.

Dart, R. A. (1925) : *Nature*, **115**, 195.

——— (1934) : *Folia Anat.*, Japan, **12**, 207.

DOOB, J. L. (1952): *Stochastic Processes*, John Wiley and Sons, New York.

FINNEY, D. J. (1956): Multivariate analysis of agricultural experiments. *Biometrics*, **12**, 67.

FISHER, R. A. (1928): The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc., A.*, **121**, 654.

——— (1939): The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugen.*, **9**, 238.

——— (1940): The precision of discriminant functions. *Ann. Eugen.*, **10**, 422.

——— (1956): *Statistical Methods and Scientific Inference*, Oliver and Boyd, London.

Galton, Francis (1889): *Natural Inheritance*, Macmillan and Co., London.

Hotelling, H. (1931): The generalization of Student's ratio. *Ann. Math. Stat.*, **2**, 360.

——— (1936): Relations between two sets of variates. *Biom.*, **28**, 321.

Hsu, P. L. (1939): On the distribution of roots of certain determinantal equations. *Ann. Eugen.*, **9**, 250.

JOHN, S. (1959): On errors in discrimination. (unpublished).

KARVE, I. and DANDEKAR, V. M. (1951): *Anthropometric Measurements of Maharashtra*, Deccan College, Monograph Series No. 8.

KEMPTHORNE, O. and NORDSKOG, A. W. (1959): Restricted selection indices. *Biometrics*, **15**, 10.

KENDALL, M. G. (1954): *Notes on Multivariate Analysis*, North Carolina University Inst. of Stat.

KERMACK, K. A. and HALDANE, J. B. S. (1950): Organic correlation and allometry. *Biom.*, **37**, 30.

KOLMOGOROFF, A. (1941): Stationary sequences in Hilbert space, (in Russian). *Bull. Math. Univ. Moscow*, **2**, 40.

LAZARSFELD, P. F. (1956): Some observations on dichotomous systems (unpublished).

LAWLEY, D. N. (1940): The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinb.*, **60**, 64.

——— (1942): Further investigations in factor estimation. *Ibid.*, **61**, 176.

LUBIN, A. (1950): Linear and non-linear discriminating functions. *British J. Psy.* (Stat. Sec.), **3**, 90.

MAHALANOBIS, P. C. (1928): A statistical study of the Chinese head. *Man in India*, **8**, 107.

——— (1930): Tests and measures of divergence. *Jour. Asiat. Soc. Bengal.*, **26**, 541.

——— (1936): On the generalized distance in statistics. *Proc. Nat. Inst. Sc. India*, **12**, 49.

——— (1941): Report on the sampling technique for forecasting the block yield of cinchona plants. (Submitted to Supt. Cinchona Culture in Bengal, in 1941).

MAHALANOBIS, P. C., MAJUMDAR, D. N. and RAO, C. R. (1949): Anthropometric survey of the United Provinces, 1941; A statistical study. *Sankhyā*, **9**, 90.

MAJUMDAR, D. N. and RAO, C. R. (1958): Bengal anthropometric survey, 1945; A statistical study. *Sankhyā*, **19**, 201. (also reprinted as a book under the title, '*Race elements of Bengal, A Quantitative Study*').

MATHER, K. (1949): *Biometrical Genetics; the Study of Continuous Variation*, Methuen, London.

MUKHERJI, R. K., RAO, C. R. and TREVOR, J. (1955): *The Ancient Inhabitants of Jebel Moya.* Cambridge University Press.

NAIR, K. R. (1952): Use of measurements of more than one biometric character regarding size of body parts for discriminating between phases in the case of six eye-striped specimens of the desert locust. *Indian Jour. of Entomology*, **14**, 126.

PANSE, V. G. (1946): An application of the discriminant function for selection in poultry. *Jour. Genetics* (London), **47**, 242.

POTI, S. J. (1948): Splitting mixed up populations. *C.S.A. Bulletin*, **1**.

RAO, C. R. (1945) : Familial correlations or the multivariate generalisation of the intraclass correlation. *Current Science*, **14**, 66.

——— (1946) : Tests with discriminant functions in multivariate analysis. *Sankhyā*, **7**, 407.

——— (1948a) : The utilization of multiple measurements in porblems of biological classification. *J. Roy. Stat. Soc.*, Series B, **10**, 159.

——— (1948b) : Tests of significance in multivariate analysis. *Biom.*, **35**, 58.

——— (1949) : On some problems arising out of discrimination with multiple characters. *Sankhyā*, **9**. 343.

——— (1950) : Statistical inference applied to classifiactory problems. *Sankhyā*, **10**, 229.

——— (1951a) : Problem of selecting individuals for various duties in a specified ratio. *Sankhyā*, **11**, 107.

——— (1951b) : The discriminant function approach in the classification of time series. *Sankhyā*, **11**, 257.

——— (1951c) : An asymptotic expansion of the distribution of Wilks' Λ criterion. *Bull. Int. Stat. Inst.*, **33** (2), 177.

——— (1952) : *Advanced Statistical Methods in Biometric Research*, John Wiley and Sons, New York.

——— (1953) : Discriminant function for genetic differentiation and selection. *Sankhyā*, **12**, 229.

——— (1954a) : A general theory of discrimination when the information about alternative populations is based on samples. *Ann. Math. Stat.*, **25**, 651.

——— (1954b) : On the use and interpretation of distance functions in statistics. *Bull. Int. Stat. Inst.*, **34**, 90.

——— (1954c) : Estimation of relative potency from multiple response data. *Biometrics*, **10**, 208.

——— (1955a) : Estimation and tests of significance in factor analysis. *Psychometrika*, **20**, 93.

——— (1955b) : Analysis of dispersion for multiply classified data with unequal numbers in cells. *Sankhyā*, **15**, 253.

——— (1956) : Analysis of dispersion with missing observations. *J. Roy. Stat. Soc.*, Series B, **18**, 259.

——— (1958) : Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1.

——— (1959) : Some problems involving linear hypotheses in multivariate analysis. *Biom.*, **46**, 49.

RAO, C. R. and SHAW, D. C. (1948) : On a formula for the prediction of cranial capacity. *Biometrics*, **4**, 247.

RAO, C. R. and SLATER, P. (1949) : Multivariate analysis applied to differences between neurotic groups. *British J. Psy.* (Stat. Sec.), **2**, 17.

ROY, J. (1951) : The distribution of certain likelihood criteria useful in multivariate analysis. *Bull. Int. Stat. Inst.*, **33**(2), 219.

ROY, J. and MITRA, S. K. (1954) : A method of selection for improvement. *Bull. Cal. Stat. Ass.*, **5**, 82.

ROY, S. N. (1939) : *p*-statistics or some generalisations in analysis of variance appropriate to multivariate problems. *Sankhyā*, **4**, 391.

TEISSIER, G. (1948) : La relation d'allometrie : sa signification statistique et biologique. *Biometrics*, **4**, 14.

WALD, A. (1947) : *Sequential Analysis*, John Wiley and Sons, New York.

WALKER, G. T. (1914) : *Mem. Ind. Met. Dept.*, **21**, 13.

WELCH, B. L. (1939) : Note on discriminant functions. *Biom.*, **31**, 218.

WILLIAMS, E. J. (1952) : Some exact tests in multivariate analysis. *Biom.*, **39**, 17.

WILKS, S. S. (1932) : Certain generalizations in the analysis of variance. *Biom.*, **24**, 471.

——— (1946) : Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. Math. Stat.*, **17**, 257.

WISHART, J. (1928) : The generalized product moment distribution in samples from a normal multivariate population. *Biom.*, **20A**, 32.

WIENER, N. and MASANI, P. (1958) : The prediction theory of multivariate stochastic processes, II. *Acta Mathematica*, **99**, 93.

WOLD, H. (1938) : *A Study in the Analysis of Time Series*, Uppsala.

YATES, F. and HEALY, M. J. R. (1951) : Statistical methods in anthropology, *Nature*. **168**, 1116.

MISCELLANEOUS

MAXIMUM LIKELIHOOD ESTIMATION FOR POSITIVELY REGULAR MARKOV CHAINS

By B. R. BHAT*

University of Western Australia, Nedlands, W.A.

SUMMARY. In this paper the author considers m. l. estimation of an unknown parameter in a simple Markov chain. The author applies ideas of Rao (1957) to Markov chains which can be approximated to a set of multinomial distributions.

1. Introduction

In a recent paper Rao (1957) has proved that for the multinomial distribution a maximum likelihood (m.l.) estimate is a root of the likelihood equation and provides a maximum of the likelihood when the parameter is restricted to the roots, with probability one, as n, the sample size tends to infinity. He has also discussed its analytical properties. In this paper we give the corresponding results for the positively regular Markov chain (MC). As it is well known that for large number of transitions, a positively regular MC behaves like a set of multinomial distributions, our result may be considered as a generalisation of Rao's work. Hence we use a notation similar to that of Rao and draw the analogy wherever possible in the course of the proof.

Let

$$\pi = (\pi_{ij}), \qquad (i, j = 1, 2, \ldots, s) \qquad \ldots \quad (1.1)$$

be the transition probability (TP) matrix of the positively regular MC with

$$\pi_i > 0, \qquad (i = 1, 2, \ldots, s)$$

as its stationary probabilities. If a realisation consists of n_{ij} transitions from state E_i to state E_j such that

$$\sum_{i,j} n_{ij} = n,$$

for large n, the logarithm of the dominant part of the likelihood function can be written as

$$L(\pi) \sim \sum_{i,j=1}^{s} n_{ij} \ln \pi_{ij} \qquad \ldots \quad (1.2)$$

(*cf.* Bartlett, 1956, § 8.2).

---

* On leave from Karnatak University, Dharwar, India. Now at the University of California, Berkeley.

Hence for long realisations from such a MC the likelihood depends only on the TP matrix π. In this paper we are concerned with the m.l. estimation of π, in the positively regular case; the corresponding results for regular but not positively regular MC will be given in a subsequent paper.

If π_{ij} depend on a certain unknown parameter θ, under certain regularity conditions to be satisfied by π_{ij}, Gani (1955) has proved that the m.l. estimator $\hat{\theta}$ of θ possesses some optimum properties. In this paper under somewhat less stringent conditions we prove that $\hat{\theta}$ makes the likelihood an absolute maximum, and derives its properties.

2. PRELIMINARY RESULTS

(a) Whittle (1955) has obtained the exact distribution of the transition numbers n_{ij}, and has shown that

$$E(n_{ij}) \rightarrow n\,\pi_i\,\pi_{ij}$$

as $n \rightarrow \infty$. Since the variance of n_{ij}/n is of order n^{-1} for large n, by the law of large numbers

$$\frac{n_{ij}}{n} \rightarrow \pi_i \pi_{ij}. \qquad \ldots \quad (2.1)$$

Similarly

$$\frac{n_{ij}}{n_{i.}} \rightarrow \pi_{ij} \qquad \ldots \quad (2.2)$$

where $n_{i.} = \sum_j n_{ij}\ (i, j = 1, 2, \ldots, s)$.

It is necessary to remark here that all the statements which we make will be true with probability tending to 1 and not in the absolute sense.

(b) A statistic is said to be consistent (Fisher Consistent (FC)), if, when it is calculated from the whole population, it is equal to the required parameter (*cf.* Kallianpur and Rao, 1955). In our case, we assume that the given realisation of n transitions is a sample from a population of infinite number of transitions. The population transition probabilities π_{ij} are thus to be viewed as the limiting values of the proportions of transitions from E_i to E_j to all possible transitions from E_i. Hence

$$\pi_{ij} = \underset{n \rightarrow \infty}{\mathrm{Lt}} \frac{n_{ij}}{n_{i.}}.$$

Defining

$$p_{ij} = \frac{n_{ij}}{n_{i.}}$$

as the sample estimate of π_{ij}, we have $T(p_{ij})$, a function of p_{ij}'s, FC for an unknown parameter θ, if

$$T(\pi_{ij}) = \theta.$$

It is to be noted that $T(p_{ij})$ need not be an unbiased estimate of θ.

Lemma 1: *If* (a) $T(p_{ij})$ *is* FC *for* θ,

(b) $T(p_{ij})$ *admits continuous first partial derivatives,*

then (i) $T(p_{ij})$ *is asymptotically normally distributed with mean* θ *and variance*

$$\text{var}\,(T) = \frac{1}{n}\sum_i \frac{1}{\pi_i}\left\{\sum_j \pi_{ij}\left(\frac{\partial T}{\partial \pi_{ij}}\right)^2 - \left(\sum_j \pi_{ij}\frac{\partial T}{\partial \pi_{ij}}\right)^2\right\} \qquad \dots \quad (2.3)$$

and (ii)

$$n\,\text{var}\,(T) \geqslant \frac{1}{i(\theta)}, \qquad \dots \quad (2.4)$$

where

$$i(\theta) = \sum_{i,j} \pi_i\pi_{ij}\left(\frac{1}{\pi_{ij}}\frac{\partial \pi_{ij}}{\partial \theta}\right)^2,$$

is the information defined by Fisher.

Proof : Expanding $T(p_{ij})$ by Taylor's series, up to the second term,

$$T(p_{ij}) = T(\pi_{ij}) + \sum_{ij}(p_{ij}-\pi_{ij})\left(\frac{\partial T}{\partial p_{ij}}\right)_{p_{ij}=\pi'_{ij}}$$

where

$$\pi'_{ij}\ \varepsilon\ (\pi_{ij},\ p_{ij}).$$

By the continuity of $\frac{\partial T}{\partial p_{ij}}$, since $\pi'_{ij} \to \pi_{ij}$ when $p_{ij} \to \pi_{ij} = E(p_{ij})$,

$$\left(\frac{\partial T}{\partial p_{ij}}\right)_{p_{ij}=\pi'_{ij}} \to \frac{\partial T}{\partial \pi_{ij}}.$$

Hence for large n, $T(p_{ij}) \to T(\pi_{ij})$ and

$$\text{var}\,(T) = \sum_{i,j}\sum_{i',j'}\frac{\partial T}{\partial \pi_{ij}}\frac{\partial T}{\partial \pi_{i'j'}}\,\text{cov}\,(p_{ij}, p_{i'j'}).$$

Since

$$\text{cov}\,(p_{ij}, p_{i'j'}) = \frac{\delta_{ii'}\,(\delta_{jj'}\pi_{ij} - \pi_{ij}\,\pi_{i'j'})}{n\pi_i}$$

where δ_{ij} is Kronecker's delta, (*cf.* Whittle, 1955),

$$\text{var}\,(T) = \frac{1}{n}\sum_i \frac{1}{\pi_i}\left\{\sum_j \pi_{ij}\left(\frac{\partial T}{\partial \pi_{ij}}\right)^2 - \left(\sum_j \pi_{ij}\frac{\partial T}{\partial \pi_{ij}}\right)^2\right\} = \frac{1}{n}\sum_i \frac{\psi_i}{\pi_i}. \qquad \dots \quad (2.5)$$

For large n the joint distribution of p_{ij} being normal (*cf.* Anderson and Goodman, 1957) T has an asymptotic normal distribution (*cf.* Rao, 1952, § 5e) with mean and variance given above. Since $T(\pi_{ij}) = \theta$

$$1 = \sum_{ij}\frac{\partial T}{\partial \pi_{ij}}\frac{\partial \pi_{ij}}{\partial \theta} = \sum_{i,j}\pi_{ij}\frac{\partial T}{\partial \pi_{ij}}\left(\frac{1}{\pi_{ij}}\frac{\partial \pi_{ij}}{\partial \theta}\right) = \sum_i \phi_i. \qquad \dots \quad (2.6)$$

Let
$$I_i(\theta) = \sum_j \pi_{ij} \left(\frac{1}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \theta}\right)^2 \quad \ldots \quad (2.7)$$

so that $i(\theta) = \Sigma\, \pi_i\, I_i(\theta)$.

Comparing ψ_i, ϕ_i and I_i, we have from Kallianpur and Rao (1955),

$$\psi_i\, I_i \geqslant \phi_i^2. \quad (i=1, 2, \ldots, s).$$

Summing up over all values of i, we have

$$\Sigma\, \psi_i\, I_i \geqslant \Sigma\, \phi_i^2.$$

Since ψ_i and I_i are positive, from Cauchy's inequality

$$\left(\sum_i \frac{\psi_i}{\pi_i}\right)(\Sigma I_i\, \pi_i) \geqslant (\sum_i |\sqrt{\psi_i\, I_i}|)^2 \geqslant (\Sigma\, \varphi_i)^2 = 1. \quad \ldots \quad (2.8)$$

Hence
$$\text{var}\,(T) \geqslant \frac{1}{n\; i(\theta)}\,.$$

So there exists a minimum variance for the FC estimator of θ, which is non-zero.

(c) Here we give a version of Rao's (1957, 1958) lemma.

Lemma 2: *If a_{ij} are positive, b_{ij} are non-negative, such that*

$$\sum_j a_{ij} = \sum_j b_{ij} \quad (i = 1, 2, \ldots, s)$$

then
$$\sum_{i,j} k_i\, a_{ij}\, ln(b_{ij}/a_{ij}) \leqslant 0 \quad (k_i \geqslant 0).$$

The equality is attained when and only when $a_{ij} = b_{ij}$ for all i and j.

Proof : From Rao's lemma

$$\sum_j a_{ij}\, ln\,(b_{ij}/a_{ij}) \leqslant 0,$$

the equality holding when and only when $a_{ij} = b_{ij}\, (j = 1, 2, \ldots, s)$. Multiplying by a non-negative constant k_i and summing up the above inequality for all values of i, we have the lemma.

As an application, we see that

$$\sum_{i,j} \frac{n_{ij}}{n_{i\cdot}} \frac{n_{i\cdot}}{n} \left[\; ln\; \pi_{ij} - ln\; \frac{n_{ij}}{n_{i\cdot}} \right] \leqslant 0,$$

if $n_{i\cdot} > 0$, the equality holding if and only if

$$\pi_{ij} = \frac{n_{ij}}{n_{i\cdot}}\, (i, j = 1, 2, \ldots, s).$$

Hence $L(\pi)$ is maximum when $n_{ij}/n_{i\cdot}$ is substituted for $\pi_{ij}\, (i, j = 1, 2, \ldots, s)$.

3. Maximum likelihood (m.l.) estimation

The proof of the following theorems proceeds in the same manner as given by Rao (1957) for the multinomial, using the results given in section 2 and therefore is not given here.

(i) If π^0 is the true TP matrix, and if the m.l. estimate $\hat{\pi}$ exists at least when $p = (p_{ij})$ belongs to the neighbourhood of π^0, $N(\pi^0)$, then $\hat{\pi}$ is a consistent estimate of π^0.

If the m.l. estimate of π^0 does not exist however close p is to π^0, then any near m.l. estimate (for definition see Rao (1957)) could be chosen.

(ii) Let π depend on a certain parameter θ, whose true value is θ_0. If θ^* is an estimator of θ which makes $L(\theta)$ a maximum, θ^* is a consistent estimator of θ, under the following condition.

Condition (A) : For any given δ,

$$\inf_{|\theta-\theta_0|>\delta} \left\{ -\sum_j \pi_{ij}(\theta_0)\, ln \left(\frac{\pi_{ij}(\theta)}{\pi_{ij}(\theta_0)} \right) \right\} = \epsilon > 0$$

for at least one i.

The above condition is analogous to assumption A_1 of Rao (1957). We will introduce three more conditions (B), (C) and (D) formulated on the lines of A_2, A_3 and A_4 respectively of Rao (1957).

(iii) Let $\pi_{ij}(\theta)$ satisfy conditions (B), (C) and (D). Then for a given δ if $p \in N(\pi^0)$, then there exists one and only one root $\hat{\theta}$ of the likelihood equation

$$\frac{dL}{d\theta} = 0$$

which differs from θ_0 by less than δ. This root, as a function of the transition frequencies is continuous at π^0 and tends to θ_0. It is also Fisher consistent.

(iv) Under conditions (A)—(D), $\hat{\theta}$ is the unique m.l. estimate at least when $p \in N(\pi^0)$.

(v) Under conditions (B), (C) and (D) $\hat{\theta}$ has asymptotic normal distribution with mean θ_0 and variance $1/n\, i(\theta_0)$, which is the minimum attainable under conditions of Lemma 1.

In fact $\hat{\theta}-\theta_0$ can be approximated by the linear function

$$-\frac{1}{i(\theta_0)} \sum_{i,j} \frac{\pi_i^0 (p_{ij}-\pi_{ij}^0)}{\pi_{ij}(\theta_0)} \cdot \frac{d\pi_{ij}^0}{d\theta_0}\,.$$

In this paper we have considered m.l. estimation of an unknown parameter in a simple Markov chain. It may be of interest to study the estimation in m-th order Markov chain when the transition probabilities depend on a number of unknown parameters.

The author is indebted to Dr. Gani for many suggestions which improved the paper considerably.

REFERENCES

ANDERSON, T. W. and GOODMAN, L. A. (1957) : Statistical inference about Markov chains. *Ann. Math. Stat.*, **28**, 89–110.

BARTLETT, M. S. (1956) : *An Introduction to Stochastic Processes.* Cambridge University Press, Cambridge.

GANI, J. (1955) : Some theorems and sufficiency conditions for the maximum likelihood estimator of an unknown parameter in a simple Markov chain. *Biometrika*, **42**, 342–359.

KALLIANPUR, G. and RAO, C. R. (1955) : On Fisher's lower bound to the asymptotic variance of an estimate. *Sankhyā*, **15**, 331–342.

RAO, C. R. (1952) : *Advanced Statistical Methods in Biometric Research.* John Wiley, New York.

——— (1957) : Maximum likelihood estimation for the multinomial distribution. *Sankhyā*, **18**, 139–148.

——— (1958) : Maximum likelihood estimation for the multinomial distribution with infinite number of cells. *Sankhyā*, **20**, 211–218.

WHITTLE, P. (1955) : Some distribution and moment formulae for the Markov chain. *J. Roy. Stat. Soc.*, Series B, **17**, 235–242.

Paper received : January, 1959.

Revised : March, 1959.

A NOTE ON THE STRUCTURE OF A CERTAIN STOCHASTIC MODEL

By S. C. DAS

Ravenshaw College, Cuttack, India

SUMMARY. Dandekar (1955) proved that the probability of a success in any $(m-1)$ consecutive trials where the process has been going on for a long time, is given by

$$P = \frac{(m-1)p}{1+(m-1)p}$$

Where p stands for the probability of a success in any trial and it is assumed that if a trial results in a success the next $(m-1)$ trials are bound to result in failures. This statement of the result is not very clear.

Basu (1955) identifies this model as a stochastic process and derives the result found by Dandekar in a slightly different form.

In this paper we identify this stochastic process which at first sight is not Markovian as a Markov chain and deduce the results found by Basu and Dandekar by an elementary method. We further find out simple expressions for the mean recurrence time and its variance in connection with this problem.

1. INTRODUCTION

Dandekar (1955) has considered a sequence of trials with the following property: if a trial results in a success, the next $(m-1)$ trials are bound to result in failures. The probability of a success in any trial in which a success is possible is p, and that of a failure is q, where $p+q = 1$. Dandekar proves by an ingenious argument that the probability of a success in any $(m-1)$ consecutive trials where the process has been going on for some time, is given by

$$P = \frac{(m-1)p}{1+(m-1)p}.$$

However, this statement of the result is not very clear.

Basu (1955) identifies this model as the stochastic process $x_0, x_1, x_2 \ldots$, characterised by the following defining postulates.

(i) Each x_i can take only the two values 0(= failure) and 1(= success) and the probability that $x_0 = 1$ is p.

(ii) For any m (or less) consecutive x_i's at most one can be 1.

(iii) If any $(m-1)$ (or more) consecutive x_i's are known to be zeros then the next x_i may be 1 with probability p.

(iv) If $x_0 = 0$, then the conditional stochastic process $x_1, x_2, \ldots$, is the same as the original process $x_0, x_1, x_2, \ldots$. Taking $P(x_0 = 1) = p$, Basu finds an expression for the probability $P_n = P\{x_n = 1\}$, and proves that in such a case $\underset{n\to\infty}{Lt}\, P_n = \frac{p}{1+(m-1)p}$, as Dandekar has already found. Basu, clarifying Dandekar's result, points out that if we do not know the x_i's preceding x_n we cannot find P_n.

In this paper we identify this stochastic process, which at first sight is not Markovian, as a Markov Chain, and deduce the results found by Basu and Dandekar by an elementary method. Observational material corresponding to this process is found, among other cases, in problems of industrial accidents. Once a worker meets with an accident he needs hospitalisation for a certain period of time for which he remains immune to further accident. For definiteness let us take a fixed period say m days as the period of hospitalisation that a worker has to undergo as soon as he meets with an accident. Now consider the states $E_0, E_1, E_2 \dots, E_m$. Let the state E_1 correspond to the situation where the worker meets with an accident and has passed the first day of his hospitalisation period. Let $E_r (r = 1, 2, 3, \dots, m-1)$ be the state corresponding to the situation when the patient has already passed r days in the hospital. E_m will be the state corresponding to recovery and will therefore be same as E_0. We will call E_0 as the state of failure and E_1 as the state of success (worker meets with an accident). Thus the state E_1 can only be reached from the state E_0. Once the state E_1 is reached, the system must necessarily go through the consecutive $m-1$ states of immunity $E_2, E_3 \dots, E_{m-1}, E_0$. Consequently any state $E_r (r \neq 0)$ can only be reached from E_{r-1} where E_0 can be reached from either E_0 or E_{m-1}. This model is a Markov Chain with the following matrix of transition probabilities.

$$\begin{array}{c|cccccc|}
 & E_0 & E_1 & E_2 & E_3 & \dots & E_{m-1} \\
E_0 & q & p & 0 & 0 & \dots & 0 \\
E_1 & 0 & 0 & 1 & 0 & 0 & 0 \\
E_2 & 0 & 0 & 0 & 1 & 0 & 0 \\
\vdots & & & & & & \\
E_{m-2} & 0 & 0 & 0 & \dots & \dots & 1 \\
E_{m-1} & 1 & 0 & 0 & \dots & \dots & 0
\end{array} = \underline{p}$$

We are here considering only one particular type of accident which when met the worker has to undergo a fixed period of hospitalisation say m days. Here each day in the factory may be considered as a trial and the day on which the accident occurs may therefore be considered as a trial resulting in a success. Excluding this day the next $(m-1)$ days in the hospital may be taken as trials resulting in failures.

But this model is a particular case of the following general model which is more closely satisfied by observational material such as is met in industrial accidents. Accidents of varying severity are usual and their periods of hospitalisation vary from accidents to accident, a fact which has been ignored by the particular model considered above.

Without any loss of generality we may retain the same notations and the same states $E_0, E_1, E_2 \dots, E_m = E_0$ as in the above particular model. There any state E_r $(r = 1, 2, \dots, m)$ corresponds to the situation where the worker meeting with an accident has remained in the hospital for r days. The state E_r therefore, consequently corresponds to the situation where the patient has to remain only $(m-r)$ days in the hospital. In the general model considered here we take any state E_r in the latter sense. Consequently E_r is either the r-th state of immunity which can be reached from the $r-1$-th state of immunity E_{r-1} with certainty or the r-th state of success (an accident of r-th type when met the worker has to remain $m-r$ days in the hospital) which can be reached from E_0, a state of failure with probability $p\alpha_{m-r}$

the suffix $m-r$ in α_{m-r} corresponding to the period of immunity, which must follow this type of accident. Thus from E_0 there is a transition of the system to any one of $E_0, E_1, \ldots, E_{m-1}$ with probabilities q, $p\alpha_{m-1}$, $p\alpha_{m-2}$, $\ldots$ $p\alpha_1$ respectively where $\sum_1^{m-1} \alpha_j = 1$. Once the state E_p $(p \neq 0)$ is reached the system must necessarily go through the states $E_{p+1}, E_{p+2} \ldots, E_m = E_0$. For this system we have the following matrix of transition probabilities.

$$\begin{array}{c|ccccccc|}
 & E_0 & E_1 & E_2 & & \ldots & E_{m-2} & E_{m-1} \\
E_0 & q & p\alpha_{m-1} & p\alpha_{m-2} & \cdots & \cdots & p\alpha_2 & p\alpha_1 \\
E_1 & 0 & 0 & 1 & 0 & \ldots & 0 & 0 \\
E_2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\vdots & & & & & & & \\
E_{m-2} & 0 & 0 & . & . & . & 0 & 1 \\
E_{m-1} & 1 & 0 & 0 & . & . & 0 & 0
\end{array} = \underline{p}$$

If we start from any known initial state E_i $(i = 0, 1, \ldots, m-1)$ the probability $p_{ij}^{(n)}$, i.e. the probability that starting from E_i we reach the state of success after n steps, is the element in the i-th row and j-th column of $\underline{p}^n$, the n-th power of the matrix $\underline{p}$. $p_{ij}^{(n)}$ will have no meaning if we do not know what the initial state E_i is; but it is possible, however, to find $\operatorname*{Lt}_{n\to\infty} p_{ij}^{(n)} = P_j$ $(j = 0, 1, \ldots, m-1)$ the stationary probabilities of the Markov Chain, which are known to be independent of the initial state E_i (Feller, 1950, p. 325).

Taking
$$\underline{P} = \left|\begin{array}{c} P_0 \\ P_1 \\ P_2 \\ \vdots \\ P_{m-1} \end{array}\right|$$

we have
$$\underline{p}'\underline{P} = \underline{P};$$

and from this we get
$$\begin{aligned}
qP_0 + P_{m-1} &= P_0 \\
\alpha_{m-1}pP_0 &= P_1 \\
\alpha_{m-2}pP_0 + P_1 &= P_2 \\
&\cdots\cdots \\
\alpha_1 pP_0 + P_{m-2} &= P_{m-1}
\end{aligned}$$

subject to the condition
$$\sum_{i=0}^{m-1} P_i = 1.$$

It follows that
$$P_0 + \alpha_{m-1}pP_0 + p(\alpha_{m-1} + \alpha_{m-2})P_0 + \ldots = 1,$$

so that
$$P_0 = \frac{1}{1 + p\{\alpha_1 + 2\alpha_2 + \ldots + (m-1)\alpha_{m-1}\}}$$

and the probability of a success of any type is

$$\sum_{i=1}^{m-1} P_i = \frac{p\{\alpha_1+2\alpha_2+\ldots+(m-1)\alpha_{m-1}\}}{1+p\{\alpha_1+2\alpha_2+\ldots+(m-1)\alpha_{m-1}\}}$$

which reduces to $\sum\limits_{i=1}^{m-1} P_i = \dfrac{(m-1)p}{1+(m-1)p}$, if we let $\alpha_1 = \alpha_2 = \ldots = \alpha_{m-2} = 0$ and $\alpha_{m-1} = 1$. This is what Dandekar gives as the probability of a success in any abrupt sequence of $(m-1)$ consecutive trials.

2. RECURRENCE TIMES

These states are all recurrent and we can, by following the method discussed in Feller (1950, pp. 347-62), find their mean recurrence times and the finite variances of these times.

Following Feller, we express the n-step transition probability of any state E_j as

$$p_{jj}^{(n)} = \frac{\rho_{jj}^{(1)}}{S_1^n}+\frac{\rho_{jj}^{(2)}}{S_2^n}+\ldots+\frac{\rho_{jj}^{(m)}}{S_m^n} = \sum_{r=1}^{m} \frac{\rho_{jj}^{(r)}}{S_r^n}, \qquad \ldots \quad (2.1)$$

where $\rho_{jj}^{(r)}$'s $(r = 1, 2, \ldots, m)$ are determinable constants and S_r's $(r = 1, 2, \ldots, m)$ are the roots of the determinantal equation

$$\begin{vmatrix} Sq-1 & \alpha_{m-1}ps & \alpha_{m-2}ps & \ldots & & & \alpha_1 ps \\ 0 & -1 & S & 0 & \ldots & & 0 \\ 0 & 0 & -1 & S & 0 & \ldots & 0 \\ \ldots & & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & & & -1 & S \\ S & 0 & 0 & \ldots & & 0 & -1 \end{vmatrix} = 0; \qquad \ldots \quad (2.2)$$

this equation can easily be shown to have no repeated roots.

On simplification (2.2) reduces to

$$p\,\{\alpha_{m-1}\,S^m+\alpha_{m-2}\,S^{m-1}+\ldots+\alpha_1 S^2\}+Sq-1 = 0. \qquad \ldots \quad (2.3)$$

It is clear that one root of (2.3) is unity and all other roots have moduli greater than unity. Let these roots be $S_1 = 1$, $S_2 \ldots S_m$; by taking the limit, (2.1) reduces to

$$\underset{n\to\infty}{\text{Lt}}\; p_{jj}^{(n)} = \rho_{jj}^{(1)}.$$

Thus the mean recurrence time μ_j for the state E_j of success is $\dfrac{1}{\rho_{jj}^{(1)}}$. The variance* σ_j^2 of this mean recurrence time is given by

$$\sigma_j^2 = \mu_j^2-\mu_j+2\mu_j^2 \sum_{r=2}^{m} \frac{\rho_{jj}^{(r)}}{S_{r-1}}. \qquad \ldots \quad (2.4)$$

---

*If the equations (5.1) and (5.3) of Feller (1950; p.362) are considered, it will be found that (5.4) reduces to our equation (2.4); Feller's result seems to be due to a printing error.

To evaluate $\rho_{jj}^{(r)}$'s we follow Feller (p. 349-50) and obtain

$$\frac{1}{\mu_j} = \rho_{jj}^{(1)} = \frac{p \sum_{k=1}^{j} \alpha_{m-k}}{1+p\left\{\sum_{k=1}^{m-1} (m-k)\alpha_{m-k}\right\}} \text{ for } j \neq 0 \qquad \dots \quad (2.5)$$

and

$$\rho_{jj}^{(r)} = \frac{p\, S_r^{m-j+1} \sum_{k=1}^{j} S_r^{j-k}\, \alpha_{m-k}}{1+p S_r \sum_{k=1}^{m-1} (m-k)\alpha_{m-k}\, S_r^{m-k}} .$$

Therefore

$$\sum_{r=2}^{n} \frac{\rho_{jj}^{(r)}}{S_{r-1}} = \sum_{r=2}^{m} \frac{p S_r^{m-j+1} \sum_{k=1}^{j} S_r^{j-k}\, \alpha_{m-k}}{\left\{1+p S_r \sum_{k=1}^{m-1} (m-k)\alpha_{m-k} S_r^{m-k}\right\}(S_{r-1})}, \qquad \dots \quad (2.6)$$

which can be calculated for any given value of m. Substituting (2.5) and (2.6) in (2.4) the variance of the recurrence time for any state $E_j (j = 1, 2 \dots, m-1)$ can be obtained.

But in the case of Dandekar's model, these results reduce to the following simpler forms.

Equation (2.3) is

$$S^m p + Sq - 1 = 0, \qquad \dots \quad (2.7)$$

which can easily be shown to have distinct roots, and we have

$$\rho_{jj}^{(1)} = \frac{p}{1+(m-1)p}, \text{ for } j = 1, 2, \dots, m-1$$

as found by Basu.

Thus the mean recurrence time for the state E_j is

$$\mu_j = \frac{1+(m-1)p}{p}. \qquad \dots \quad (2.8)$$

Similarly, we have that

$$\rho_{jj}^{(r)} = \frac{S_r^m p}{1-(m+1)S_r^m p}$$

which by (2.7) reduces to

$$\rho_{jj}^{(r)} = \frac{1-S_r q}{(m-1)q(k-S_r)}$$

where

$$K = \frac{m}{(m-1)q} .$$

It follows that

$$\frac{\rho_{jj}^{(r)}}{S_r - 1} = -\frac{1}{(m-1)(k-1)q}\left\{\frac{kq-1}{k-S_r} + \frac{p}{1-S_r}\right\}$$

and the sum

$$\sum_{r=2}^{m} \frac{\rho_{jj}^{(r)}}{S_r-1} = -\frac{kq-1}{(m-1)(k-1)q}\left\{\sum_{r=2}^{m}\frac{1}{k-S_r}\right\}-\frac{p}{(m-1(k-1)q}\left\{\sum_{r=2}^{m}\frac{1}{(1-S_r)}\right\} \quad \ldots \quad (2.9)$$

By using the relation that the sum of the products of the roots $S_2, S_3, \ldots, S_m$ of the equation (2.7) taken r $(r < m-1)$ at a time is equal to $(-1)^r$ and $S_2, S_3, \ldots S_m, = (-1)^m/p$, (2.9) reduces to

$$\sum_{r=2}^{m} \frac{\rho_{jj}^{(r)}}{S_r-1} = -\frac{kq-1}{(m-1)(k-1)q}\left\{\frac{(m-1)k^{m-2}+(m-2)k^{m-3}+\ldots+2k+1}{k^{m-1}+k^{m-2}+\ldots+k+1/p}\right\}-$$

$$-\frac{p}{2(m-1)(k-1)q}\left\{\frac{m(m-1)p}{1+(m-1)p}\right\}$$

$$= -\frac{(m-1)q}{p\mu_1^2}\left\{\frac{(m-1)k^m-mk^{m-1}+1}{(m-1)k^mp+1}\right\}-\frac{m(m-1)}{2\mu_1^2}$$

$$= -\frac{(m-1)q}{p\mu_1^2}-\frac{m(m-1)}{2\mu_1^2}. \quad \ldots \quad (2.10)$$

The variance of the mean recurrence time in this case is found to be

$$\sigma_j^2 = q/p^2, \quad \text{for} \quad j = 1, 2, \ldots (m-1)$$

a remarkably simple result.

The mean recurrence time and its variance for the state E_0 are given by

$$\mu_0 = 1+(m-1)p$$

and

$$\sigma_0^2 = (m-1)^2qp.$$

References

Basu, D. (1955) : A note on the structure of a stochastic model considered by V. M. Dandekar. *Sankhyā*, **15**, 251–52.

Dandekar, V. M. (1955) : Certain modified forms of Binomial and Poisson distributions. *Sankhyā*, **15**, 237–50.

Feller, W. (1950) : *An Introduction to Probability Theory and its Applications.* John Wiley, New York.

Paper received : February, 1956.
Revised : June, 1959.

A SHORT TABLE OF THE HYPERGEOMETRIC FUNCTION ${}_0F_1(c;x)$†

By J. EDWARD JACKSON*

Virginia Polytechnic Institute USA

SUMMARY. The generalized hypergeometric function ${}_0F_1(c;x)$ has certain statistical applications particularly with regard to sequential multivariate techniques. This function is tabulated for $x = .1(.1)1(1)10(10)100(50)1000$ and $c = .5(.5)5$.

The function

$$ {}_0F_1(c;\, x) = 1 + \frac{x}{c} + \frac{x^2}{c(c+1)2!} + \frac{x^3}{c(c+1)(c+2)3!} + \dots $$

is a special case of the generalized hypergeometric function and is a solution of the differential equation : $xy''+cy'-y = 0$ (Erdelyi, *et al.*, 1953). It also has applications in the field of mathematical statistics. In particular, the non-central χ^2-distribution with p degrees of freedom and noncentrality parameter $n\lambda^2$ can be written in the form :

$$ f(\chi^2) = \frac{e^{-(\chi^2+n\lambda^2)/2}}{2^{(p/2)}} \, \frac{(\chi^2)^{(p-2)/2}}{\Gamma(p/2)} \, {}_0F_1(p/2;\, n\lambda^2\chi^2/4). $$

Some multivariate sequential tests for means (Jackson and Bradley, 1959) involve this function as well as the confluent hypergeometric function ${}_1F_1(a, c;\, x)$ which has already been tabulated in this Journal (Nath, 1952; Rushton, 1954; Rushton and Lang, 1954). The function ${}_0F_1(c;\, x)$ is covergent for all finite x.

One of the reasons that this function has not received much attention in the literature is that it is related to several other well known transcendental functions, among them:

(a) Modified Bessel Function of the first kind.

$$ {}_0F_1(c;\, x) = \Gamma(c)x^{\frac{1-c}{2}} I_{c-1}(2\sqrt{x}). $$

(b) Confluent hypergeometric function.

$$ {}_0F_1(c;\, x) = e^{-2\sqrt{x}} \, {}_1F_1(c-\tfrac{1}{2},\, 2c-1;\, 4\sqrt{x}). $$

(c) Whittaker's M-function.

$$ {}_0F_1(c;\, x) = (4\sqrt{x})^{\frac{1-c}{2}} M_{0,\, c-1}(4\sqrt{x}). $$

However, for many applications, these alternative functions are not tabulated to the extent that is required and, in addition, the use of these functions makes inverse solutions more difficult since the argument x is involved twice in each of these expressions. For these reasons, it seemed advisable to prepare a table of the function ${}_0F_1(c;\, x)$ itself.

---

† Sponsored by the Office of Naval Research, Department of the Navy; Contract Number: NONR-2352(01). Task Order 042-019 with the Virginia Polytechnic Institute. Reproduction in whole or in part is permitted for any purpose of the United States Government.

* Now with the Eastman Kodak Company, N. Y., USA.

The present table gives the value of the function ${}_0F_1(c; x)$ to six significant figures for $x = .1(.1)1(1)10(10)100(50)1000$ and $c = .5(.5)5$. Thus the table has a uniform graduation for c and is divided into four "bands" for values of x. The value in parentheses directly following each tabulated value indicates the power of 10 by which the tabulated value should be multiplied to produce the true result. For example :

$${}_0F_1(4.5; 100) = 1.17108(5) = 117108.$$

When interpolating, it is best to work with $\log {}_0F_1(c; x)$ rather than the function itself. In general, when interpolating for given values of x within "bands", four-point Lagrangian interpolation will yield four-digit accuracy. For small values of x, linear interpolation (of the logs) will suffice. When interpolating near either end of a band, four-point Newton interpolation formulae will prove quite satisfactory. When interpolating for given values of c, the results will not be as accurate. If ${}_0F_1(c; x)$ is to be obtained when neither c nor x appear in the table, one must resort either to double interpolation or to a numerical solution of the original differential equation itself.

The author wishes to acknowledge the assistance of Professor R. J. Freund in the computation of the tabulated results.

REFERENCES

ERDELYI, A., MAGNUS, W., OBERHETTINGER, F. and TRICOMI, F. G. (1953): *Higher Transcendental Functions*, McGraw-Hill, New York, Sections 4.2 and 7.2.2.

JACKSON, J. E., and BRADLEY, R. A. (1959): Multivariate sequential procedures for testing means. *The Development of Statistical Methods for Experimental Designs in Quality Control and Surveillance Testing*, Technical Report No. 10, Virginia Polytechnic Institute Blacksburg, Virginia, U.S.A.

NATH, P. (1951): Confluent hypergeometric function. *Sankhyā*, **11**, 155–166.

RUSHTON, S. (1954): On the confluent hypergeometric function $M(\alpha, \gamma; x)$. *Sankhyā*, **13**, 369–376.

RUSHTON, S. and LANG, E. D. (1954): Tables of the confluent hypergeometric function. *Sankhyā*, **13**, 377–411.

Paper received : August, 1959.

TABLE OF ${}_0F_1(c; x)$

	x			
c	.1	.2	.3	.4
.5	1.20676 (0)	1.42739 (0)	1.66245 (0)	1.91252 (0)
1.0	1.10253 (0)	1.21022 (0)	1.32326 (0)	1.44182 (0)
1.5	1.06801 (0)	1.13877 (0)	1.21235 (0)	1.28883 (0)
2.0	1.05084 (0)	1.10339 (0)	1.15769 (0)	1.21379 (0)
2.5	1.04058 (0)	1.08232 (0)	1.12526 (0)	1.16942 (0)
3.0	1.03375 (0)	1.06836 (0)	1.10383 (0)	1.14018 (0)
3.5	1.02889 (0)	1.05843 (0)	1.08862 (0)	1.11949 (0)
4.0	1.02525 (0)	1.05101 (0)	1.07729 (0)	1.10409 (0)
4.5	1.02243 (0)	1.04526 (0)	1.06851 (0)	1.09219 (0)
5.0	1.02017 (0)	1.04067 (0)	1.06152 (0)	1.08272 (0)

	x			
c	.5	.6	.7	.8
.5	2.17818 (0)	2.46004 (0)	2.75873 (0)	3.07487 (0)
1.0	1.56608 (0)	1.69623 (0)	1.83246 (0)	1.97496 (0)
1.5	1.36830 (0)	1.45084 (0)	1.53653 (0)	1.62547 (0)
2.0	1.27172 (0)	1.33155 (0)	1.39330 (0)	1.45703 (0)
2.5	1.21483 (0)	1.26151 (0)	1.30950 (0)	1.35882 (0)
3.0	1.17744 (0)	1.21562 (0)	1.25473 (0)	1.29480 (0)
3.5	1.15104 (0)	1.18328 (0)	1.21623 (0)	1.24990 (0)
4.0	1.13143 (0)	1.15931 (0)	1.18774 (0)	1.21673 (0)
4.5	1.11629 (0)	1.14083 (0)	1.16582 (0)	1.19125 (0)
5.0	1.10427 (0)	1.12617 (0)	1.14844 (0)	1.17108 (0)

	x			
c	.9	1	2	3
.5	3.40914 (0)	3.76220 (0)	8.48897 (0)	1.59895 (1)
1.0	2.12393 (0)	2.27959 (0)	4.25235 (0)	7.15900 (0)
1.5	1.71774 (0)	1.81343 (0)	2.98041 (0)	4.60674 (0)
2.0	1.52280 (0)	1.59064 (0)	2.39483 (0)	3.46865 (0)
2.5	1.40950 (0)	1.46157 (0)	2.06571 (0)	2.84570 (0)
3.0	1.33585 (0)	1.37790 (0)	1.85752 (0)	2.46023 (0)
3.5	1.28431 (0)	1.31946 (0)	1.71505 (0)	2.20131 (0)
4.0	1.24630 (0)	1.27644 (0)	1.61195 (0)	2.01684 (0)
4.5	1.21714 (0)	1.24350 (0)	1.53412 (0)	1.87946 (0)
5.0	1.19410 (0)	1.21749 (0)	1.47343 (0)	1.77358 (0)

TABLE OF ${}_0F_1(c; x)$—*continued*

c	x = 4	5	6	7
.5	2.73082 (1)	4.37775 (1)	6.70801 (1)	9.93234 (1)
1.0	1.13019 (1)	1.70578 (1)	2.48921 (1)	3.53766 (1)
1.5	6.82248 (0)	9.78638 (0)	1.36911 (1)	1.87694 (1)
2.0	4.87973 (0)	6.70929 (0)	9.05436 (0)	1.20304 (1)
2.5	3.84108 (0)	5.09866 (0)	6.67362 (0)	8.63078 (0)
3.0	3.21109 (0)	4.13939 (0)	5.27926 (0)	6.67034 (0)
3.5	2.79506 (0)	3.51579 (0)	4.38596 (0)	5.43140 (0)
4.0	2.50296 (0)	3.08387 (0)	3.77510 (0)	4.59432 (0)
4.5	2.28816 (0)	2.77002 (0)	3.33618 (0)	3.99922 (0)
5.0	2.12441 (0)	2.53325 (0)	3.00832 (0)	3.55889 (0)

c	x = 8	9	10	20
.5	1.43125 (2)	2.01716 (2)	2.79056 (2)	3.83193 (3)
1.0	4.92086 (1)	6.72344 (1)	9.04759 (1)	1.03763 (3)
1.5	2.53006 (1)	3.36189 (1)	4.41223 (1)	4.28423 (2)
2.0	1.57742 (1)	2.04473 (1)	2.62399 (1)	2.18638 (2)
2.5	1.10460 (1)	1.40081 (1)	1.76200 (1)	1.27632 (2)
3.0	8.35859 (0)	1.03971 (1)	1.28472 (1)	8.18993 (1)
3.5	6.68180 (0)	8.17116 (0)	9.93836 (0)	5.63984 (1)
4.0	5.56170 (0)	6.70012 (0)	8.03558 (0)	4.10217 (1)
4.5	4.77340 (0)	5.67476 (0)	6.72144 (0)	3.11645 (1)
5.0	4.19534 (0)	4.92935 (0)	5.77396 (0)	2.45265 (1)

c	x = 30	40	50	60
.5	2.86041 (4)	1.55743 (5)	6.93141 (5)	2.67320 (6)
1.0	6.97878 (3)	3.53018 (4)	1.48419 (5)	5.46442 (5)
1.5	2.61118 (3)	1.23126 (4)	4.90124 (4)	1.72554 (5)
2.0	1.21452 (3)	5.35631 (3)	2.02334 (4)	6.82291 (4)
2.5	6.49822 (2)	2.68932 (3)	9.66192 (3)	3.12580 (4)
3.0	3.84284 (2)	1.49727 (3)	5.12743 (3)	1.59404 (4)
3.5	2.45170 (2)	9.02180 (2)	2.95129 (3)	8.83101 (3)
4.0	1.66047 (2)	5.78855 (2)	1.81271 (3)	5.22887 (3)
4.5	1.18023 (2)	3.90937 (2)	1.17436 (3)	3.27061 (3)
5.0	8.72949 (1)	2.75525 (2)	7.95532 (2)	2.14231 (3)

TABLE OF $_0F_1(c; x)$—*continued*

	x			
c	70	80	90	100
.5	9.24925 (6)	2.93674 (7)	8.69217 (7)	2.42582 (8)
1.0	1.81805 (6)	5.58012 (6)	1.60300 (7)	4.35583 (7)
1.5	5.52748 (5)	1.64169 (6)	4.58118 (6)	1.21291 (7)
2.0	2.10702 (5)	6.06180 (5)	1.64456 (6)	4.24550 (6)
2.5	9.31768 (4)	2.59929 (5)	6.86171 (5)	1.72840 (6)
3.0	4.59241 (4)	1.24348 (5)	3.19677 (5)	7.86255 (5)
3.5	2.46199 (4)	6.47700 (4)	1.62292 (5)	3.90027 (5)
4.0	1.41238 (4)	3.61373 (4)	8.83258 (4)	2.07554 (5)
4.5	8.56961 (3)	2.13455 (4)	5.09327 (4)	1.17108 (5)
5.0	5.45149 (3)	1.32317 (4)	3.08468 (4)	6.94441 (4)

	x			
c	150	200	250	300
.5	2.17254 (10)	9.60888 (11)	2.70749 (13)	5.53824 (14)
1.0	3.52074 (9)	1.44809 (11)	3.85702 (12)	7.53541 (13)
1.5	8.86938 (8)	3.39725 (10)	8.56184 (11)	1.59875 (13)
2.0	2.81537 (8)	1.00569 (10)	2.40051 (11)	4.28731 (12)
2.5	1.04192 (8)	3.47593 (9)	7.86562 (10)	1.34459 (12)
3.0	4.31894 (7)	1.34752 (9)	2.89358 (10)	4.73779 (11)
3.5	1.95686 (7)	5.71811 (8)	1.16629 (10)	1.83037 (11)
4.0	9.53391 (6)	2.61280 (8)	5.06676 (9)	7.62707 (10)
4.5	4.93639 (6)	1.27055 (8)	2.34476 (9)	3.38787 (10)
5.0	2.69244 (6)	6.51744 (7)	1.14571 (9)	1.59003 (10)

	x			
c	350	400	450	500
.5	8.88753 (15)	1.17693 (17)	1.33206 (18)	1.32193 (19)
1.0	1.16322 (15)	1.48948 (16)	1.63659 (17)	1.58168 (18)
1.5	2.37529 (14)	2.94231 (15)	3.13970 (16)	2.95593 (17)
2.0	6.13399 (13)	7.35370 (14)	7.62351 (15)	6.99397 (16)
2.5	1.85357 (13)	2.15157 (14)	2.16777 (15)	1.93856 (16)
3.0	6.29643 (12)	7.07970 (13)	6.93493 (14)	6.04698 (15)
3.5	2.34636 (12)	2.55671 (13)	2.43577 (14)	2.07156 (15)
4.0	9.43603 (11)	9.96859 (12)	9.24002 (13)	7.66713 (14)
4.5	4.04734 (11)	4.14727 (12)	3.74149 (13)	3.02996 (14)
5.0	1.83526 (11)	1.82485 (12)	1.60291 (13)	1.26726 (14)

TABLE OF ${}_0F_1(c; x)$—*continued*

			x		
c	550	600	650	700	750
.5	1.17270 (20)	9.43990 (20)	6.97799 (21)	4.78270 (22)	3.06378 (23)
1.0	1.36991 (19)	1.07888 (20)	7.81632 (20)	5.25848 (21)	3.31069 (22)
1.5	2.50020 (18)	1.92691 (19)	1.36849 (20)	9.03846 (20)	5.59368 (21)
2.0	5.77869 (17)	4.35933 (18)	3.03560 (19)	1.96865 (20)	1.19780 (21)
2.5	1.56504 (17)	1.15590 (18)	7.89362 (18)	5.02749 (19)	3.00785 (20)
3.0	4.77134 (16)	3.45096 (17)	2.31162 (18)	1.44617 (19)	8.50909 (19)
3.5	1.59798 (16)	1.13207 (17)	7.43976 (17)	4.57270 (18)	2.64645 (19)
4.0	5.78352 (15)	4.01424 (16)	2.58871 (17)	1.56345 (18)	8.90172 (18)
4.5	2.23561 (15)	1.52059 (16)	9.62452 (16)	5.71277 (17)	3.20040 (18)
5.0	9.14835 (14)	6.09908 (15)	3.78969 (16)	2.21114 (17)	1.21903 (18)

			x		
c	800	850	900	950	1000
.5	1.84661 (24)	1.05311 (25)	5.71003 (25)	2.95583 (26)	1.46610 (27)
1.0	1.96334 (23)	1.10276 (24)	5.89407 (24)	3.00997 (25)	1.47385 (26)
1.5	3.26438 (22)	1.80606 (23)	9.51672 (23)	4.79499 (24)	2.31811 (25)
2.0	6.87984 (21)	3.74986 (22)	1.94825 (23)	9.68609 (23)	4.62374 (24)
2.5	1.70059 (21)	9.13275 (21)	4.67905 (22)	2.29569 (23)	1.08219 (24)
3.0	4.73636 (20)	2.50650 (21)	1.26650 (22)	6.13285 (22)	2.85523 (23)
3.5	1.45046 (20)	7.56500 (20)	3.77034 (21)	1.80214 (22)	8.28710 (22)
4.0	4.80465 (19)	2.47003 (20)	1.21440 (21)	5.73019 (21)	2.60293 (22)
4.5	1.70138 (19)	8.62261 (19)	4.18252 (20)	1.94847 (21)	8.74405 (21)
5.0	6.38385 (18)	3.18988 (19)	1.52675 (20)	7.02295 (20)	3.11393 (21)

APPROXIMATE CONFIDENCE INTERVAL FOR LINEAR FUNCTIONS OF MEANS OF *K* POPULATIONS WHEN THE POPULATION VARIANCES ARE NOT EQUAL

By SAIBAL KUMAR BANERJEE
Indian Statistical Institute, Calcutta

SUMMARY. It is shown that given k-samples of n_i units from populations $N_i\ (m_i,\ \sigma_i^2)$ $(i=1,2..,k)$ a confidence interval for any linear function $\sum_1^k c_i m_i$ of population means with confidence coefficient not less than any pre-assigned probability α is possible in terms of sample estimates of population means and variances and tabulated values of Student's t-table.

Given k samples of n_i $(i = 1, 2, ..., k)$ units from k normal populations $N_i(m_i, \sigma_i^2)$, with sample estimates of population means and variances $\bar{x}_i$ and $s_i^2 (i = 1, 2, ..., k)$ a confidence interval for any linear function of the population means with approximate confidence co-efficient is possible. Briefly the method is indicated for the case of two samples.

Let P denote the probability of the event

$$(\bar{x}_1+\bar{x}_2-m_1-m_2)^2 \leqslant \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} \qquad \dots \quad (1)$$

where t_1 and t_2 correspond to t-values of Student's t-table satisfying the relation

$$\frac{1}{\sqrt{\nu_i}} \frac{1}{B\left(\frac{\nu_i}{2}, \frac{1}{2}\right)} \int_{-t_i}^{t_i} \left(1+\frac{t^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}} dt = \alpha \qquad \dots \quad (2)$$

$$(\nu_i = n_i - 1; \quad i = 1, 2)$$

Since for fixed s_1^2 and s_2^2

$$\frac{(\bar{x}_1+\bar{x}_2-m_1-m_2)^2}{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \qquad \dots \quad (3)$$

is distributed as a χ^2 variate with 1 d.f.,

$$P = \int_0^\infty \int_0^\infty f_1(s_1^2, \sigma_1^2, n_1) f_2(s_2^2, \sigma_2^2, n_2)\, g(s_1^2, s_2^2) ds_1^2 ds_2^2 \qquad \dots \quad (4)$$

where

$$f_1(s_1^2, \sigma_1^2, n_1) = \text{frequency function of } s_1^2$$
$$f_2(s_2^2, \sigma_2^2, n_2) = \text{frequency function of } s_2^2$$

and

$$g(s_1^2, s_2^2) = \int_0^{\frac{t_1^2 s_1^2/n_1 + t_2^2 s_2^2/n_2}{\sigma_1^2/n_1+\sigma_2^2/n_2}} \frac{1}{2} \cdot \frac{1}{\Gamma(\frac{1}{2})} \cdot e^{-\frac{\chi^2}{2}} \left(\frac{\chi^2}{2}\right)^{-\frac{1}{2}} d\chi^2.$$

Now $\int_0^z e^{-y} y^{-\frac{1}{2}}\, dy$ is an upward convex function of z and therefore

$$g(s_1^2, s_2^2) \geqslant \omega_1 \int_0^{t_1^2 s_1^2/\sigma_1^2} f(\chi^2) d\chi^2 + \omega_2 \int_0^{t_2^2 s_2^2/\sigma_2^2} f(\chi^2) d\chi^2 \qquad \dots \quad (5)$$

where

$$f(\chi^2) = \tfrac{1}{2}\,\frac{1}{\Gamma(\frac{1}{2})}\cdot e^{-\frac{\chi^2}{2}}\left(\frac{\chi^2}{2}\right)^{-\frac{1}{2}}$$

and

$$\omega_i = \frac{\sigma_i^2/n_i}{\sigma_1^2/n_1+\sigma_2^2/n_2}\,, \qquad (i = 1, 2).$$

As

$$\int_0^{\infty}\int_0^{\infty} f_1(s_1^2, \sigma_1^2, n_1)f_2(s_2^2, \sigma_2^2, n_2)\left\{\int_0^{t_i^2s_i^2/\sigma_i^2} f(\chi^2)d\chi^2\right\} ds_1^2ds_2^2 = \alpha \qquad \ldots \quad (6)$$

$$(i = 1, 2)$$

$$\text{Prob}\left\{(\bar{x}_1+\bar{x}_2-m_1-m_2)^2 \leqslant \frac{t_1^2s_1^2}{n_1}+\frac{t_2^2s_2^2}{n_2}\right\} \geqslant \omega_1\alpha+\omega_2\alpha = \alpha \qquad \ldots \quad (7)$$

so that

$$\text{Prob}\left\{\bar{x}_1+\bar{x}_2-\sqrt{\sum_1^2\frac{t_i^2s_i^2}{n_i}} \leqslant m_1+m_2 \leqslant \bar{x}_1+\bar{x}_2+\sqrt{\sum_1^2\frac{t_i^2s_i^2}{n_i}}\right\} \geqslant \alpha. \qquad \ldots \quad (8)$$

Also it can be readily shown that if c_1 and c_2 are known constants

$$\text{Prob}\left\{\sum_1^2 c_i\bar{x}_i-\sqrt{\sum_1^2\frac{t_i^2c_i^2s_i^2}{n_i}} \leqslant \sum_1^2 c_im_i \leqslant \sum_1^2 c_i\bar{x}_i+\sqrt{\sum_1^2\frac{t_i^2c_i^2s_i^2}{n_i}}\right\} \geqslant \alpha. \qquad \ldots \quad (9)$$

Further extending to k populations it can be shown that

$$\text{Prob}\left\{\sum_1^k c_i\bar{x}_i-\sqrt{\sum_1^k\frac{t_i^2c_i^2s_i^2}{n_i}} \leqslant \sum_1^k c_im_i \leqslant \sum_1^k c_i\bar{x}_i+\sqrt{\sum_1^k\frac{t_i^2c_i^2s_i^2}{n_i}}\right\} \geqslant \alpha. \qquad (10)$$

For the case of two populations if $c_1 = 1$ and $c_2 = -1$ in (9) the following relation is established :

$$\text{Prob}\left\{|\bar{x}_1-\bar{x}_2-(m_1-m_2)| \leqslant \sqrt{\frac{t_1^2s_1^2}{n_1}+\frac{t_2^2s_2^2}{n_2}}\right\} \geqslant \alpha$$

which is the two-means problem. Cochran and Cox at page 92 of the book "Experimental Designs" (1950 edition) suggest an approximate result for this case which in the notation of the present paper is

$$|\bar{x}_1-\bar{x}_2-(m_1-m_2)| \leqslant \left(\frac{t_1s_1^2}{n_1}+\frac{t_2s_2^2}{n_2}\right)\Big/\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\,.$$

Since

$$\left(\frac{t_1s_1^2}{n_1}+\frac{t_2s_2^2}{n_2}\right)^2 \leqslant \left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)\left(\frac{t_1^2s_1^2}{n_1}+\frac{t_2^2s_2^2}{n_2}\right),$$

if $n_1 \neq n_2$ this result gives a stronger result than the present result. To the best of the knowledge of the author he has not seen the result suggested as approximation proved in any published literature nor the approximate nature of the result spelt out. The approximate nature of the present result is to the effect that for all values of (σ_1^2, σ_2^2) or $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$ the true value will be covered with probability not less than pre-assigned probability α.

REFERENCES

COURANT, R. (1957): *Differential and Integral Calculus*, Volume II, Blackie and Son Limited, London and Glasgow.

COX and COCHRAN (1950): *Experimental Designs*, Wiley Publication in Statistics, New York.

Paper received : October, 1958.

Paper revised : March, 1960.

MEASUREMENT OF ATTRITION IN TECHNICAL EDUCATION

By K. C. SEAL*

Planning Commission, India

SUMMARY. The existing shortage of technical manpower can be partially met by reducing the wastage during the training period of the students admitted for technical training. For proper assessment of attrition during the course of training a detailed cohort study of each student throughout his training period is really required. Such a study for every training institution in the country (unless the total number of such institutions are very few) will be a formidable task since, mere collection of such detailed statistics for a number of years about each student admitted in any institution, will be extremely difficult and time-consuming. It is indicated in the present paper how under some realistic model certain broad conclusions regarding wastage can be arrived at from the usually available data for most of the institutions such as (i) total number of admissions each year, (ii) total number of students appearing at the qualifying examination every year, and (iii) total number of students qualifying every year. As for instance, it is shown from an analysis of the available material on (i) and (ii) only, that for medical graduates about 90% of the students admitted during 1947–50, qualified in about 7 years' time instead of the usual duration of about 5 years.

1. INTRODUCTION

For planning purposes, one of the most important problems deserving immediate attention is how to ensure adequate supply of technical manpower that may be required for the anticipated progress in future years. Improvement of training facilities in the form of opening up of new institutions and also increasing the number of admissions per institution through some planned development by stages might be immediately called for to overcome the present shortage of technical manpower available in the country. This shortage of technical manpower can also be partially met by reducing the wastage during the training period of the students. For this purpose, a study of all the factors contributing to wastage during training will evidently be required and suitable steps should be taken to reduce at least, if it is not possible to eliminate completely, the effects of these factors as far as possible. Importance of such a study of the wastage problem will depend naturally on the existing magnitude of the problem—those types of technical education in which the magnitude of wastage is considerable need only be considered to begin with. Measurements of such attrition in technical education is also required for getting a reliable estimate of the technical manpower that will be available in future years.

For proper assessment of all types of wastage during the course of training such as (i) number of temporary or permanent drop-outs before the qualifying examination, (ii) number of failures together with their frequency of attempts at the qualifying examination, and so forth, a detailed cohort study of each student admitted in the training course throughout his training period is really required. Such a study for every training institution in the country will be a formidable task since mere collection of such detailed data for a number of years about each student admitted in any institution will be extremely difficult and time-consuming. Such cohort study can, of course, be taken up for a few selected institutions in the country but measures of attrition as may be estimated from such a sample of students will not usually be valid for the entire country. Hence the problem may be somewhat modified as follows :

---

* Now with Labour Bureau, Simla.

From the data at present available for most of the institutions such as :

(i) total number of admissions each year

(ii) total number of students appearing at the qualifying examination every year; and

(iii) total number of students qualifying every year,

is it possible to get some idea of the wastage during the training period? Clearly from such limited information it will not be possible to provide estimates of various types of attrition. The purpose of the present paper is to indicate how certain broad conclusions regarding wastage can be arrived at from such limited data. It may also be noted here that although attrition during the training course is only considered in this paper, yet it is quite plausible that for certain categories of technical personnel, such as nurses, attrition after completion of the training period, as for instance, due to marriage of nurses, might be more important.

2. PROBLEM

Suppose in a particular year t, A_t students are admitted for technical training and that after a specific period of d years which may be at the time of appearing at the qualifying examination or just after the results of the qualifying examination are known, Q_{t+d} students are found to be present. Clearly Q_{t+d} comprises a portion p_{0t} of all the regular students A_t admitted d years earlier and certain proportions, say p_{it} of A_{t-i}, $i = 1, \ldots, r-1$, students admitted $d+i$ years earlier. Thus

$$Q_{t+d} = \sum_{i=0}^{r-1} p_{it} A_{t-i}. \qquad \ldots \quad (2.1)$$

Although p_{it} is expected to vary with t, it is assumed that there exists some parameter p_i which can perhaps be approximated by taking the average of p_{it}'s over a sufficient length of time t till a stable value of the average is attained and that

$$p_{it} = p_i + \epsilon_{it}, \qquad \ldots \quad (2.2)$$

where ϵ_{it} is a random variable with expectation in respect of t zero and variance inversely proportional to A_{t-i}, namely

$$\text{var}\ (\epsilon_{it}) = \sigma_i^2 / A_{t-i}. \qquad \ldots \quad (2.3)$$

Thus with a large A_{t-i} the variation of the error component ϵ_{it}, if we may say so, is likely to be small. Clearly the minimum variance linear unbiased estimate of p_i will be given by the weighted average

$$\hat{p}_i = \sum_t p_{it} A_{t-i} / \sum_t A_{t-i}, \qquad \ldots \quad (2.4)$$

where $\sum_t$ denotes summation over the number of years for which the data are available. It is readily verified that

$$\text{var}\ (\hat{p}_i) = \sigma_i^2 \Big/ \sum_t A_{t-i}. \qquad \ldots \quad (2.5)$$

Since p_{it}'s are unknown, it is not possible for us to estimate $\hat{p}_i$ from equation (2.4) and some alternative method has to be adopted for estimating p_i's. It may be noted here that although it is certainly desirable to have estimates of each p_i, $i = 0, 1..., r-1$, in case it is found to be not easily possible our minimum requirement is to get an estimate of $1 - \sum_{i=0}^{r-1} p_i$, i.e. of the total percentage of wastage during the entire training period for a group of students whose percentage of passes in several consecutive attempts is unaffected by the error component ϵ_i's.

The following method is suggested for estimating at least the total wastage. From equation (2.1) summing over successive values of t from t to $t+n-1$, we get

$$\sum_t^{t+n-1} Q_{t+d} = \sum_t^{t+n-1} \sum_{i=0}^{r-1} p_{it} A_{t-i} = \sum_{i=0}^{r-1} \sum_t^{t+n-1} p_{it} A_{t-i} = \sum_{i=0}^{r-1} \Big(\sum_t^{t+n-1} A_{t-i} \Big) \tilde{p}_i, \quad \dots \ (2.6)$$

where

$$\tilde{p}_i = \sum_t^{t+n-1} p_{it} A_{t-i} \Big/ \sum_t^{t+n-1} A_{t-i} \quad \dots \ (2.7)$$

so that when n coincides with the number of years for which the data are available, $\tilde{p}_i$ becomes identical with the best estimate $\hat{p}_i$ of p_i. It is easily seen that

$$\text{var } (\tilde{p}_i) = \sigma_i^2 \Big/ \sum_t^{t+n-1} A_{t-i} \quad \dots \ (2.8)$$

and for moderately large n and A_{t-i}'s, var $(\tilde{p}_i)$ will be very small, so that in such a situation $\tilde{p}_i$ is expected to be very close to $\hat{p}_i$.

From equation (2.6) dividing both the sides by $\sum_t^{t+n-1} A_t$, we get

$$\frac{\sum_t^{t+n-1} Q_{t+d}}{\sum_t^{t+n-1} A_t} = \sum_{i=0}^{r-1} \left(\frac{\sum_t^{t+n-1} A_{t-i}}{\sum_t^{t+n-1} A_t} \right) \tilde{p}_i \quad \dots \ (2.9)$$

and

$$\text{var} \left(\frac{\sum_t^{t+n-1} Q_{t+d}}{\sum_t^{t+n-1} A_t} \right) = \sum_{i=0}^{r-1} \frac{\sum_t^{t+n-1} A_{t-i}}{\left(\sum_t^{t+n-1} A_t \right)^2} \sigma_i^2, \quad \dots \ (2.10)$$

since it is expected that ϵ_{it} and $\epsilon_{i't'}$ for $i \neq i'$ will be uncorrelated for any t and t'.

Under the simplifying assumption of

$$A_{t-i} = \alpha_i A_t, \qquad \ldots \quad (2.11)$$

the equations (2.9) and (2.10) reduce to

$$\sum_{t}^{t+n-1} Q_{t+d} \Big/ \sum_{t}^{t+n-1} A_t = \sum_{i=0}^{r-1} \alpha_i \tilde{p}_i, \qquad \ldots \quad (2.12)$$

$$\text{var}\left(\sum_{t}^{t+n-1} Q_{t+d} \Big/ \sum_{t}^{t+n-1} A_t \right) = \sum_{i=0}^{r-1} \frac{\alpha_i \sigma_i^2}{\sum\limits_{t}^{t+n-1} A_t} = \frac{r}{n} \frac{\bar{\sigma}^2}{\bar{A}_t} \qquad \ldots \quad (2.13)$$

where

$$\bar{\sigma}^2 = \sum_{i=0}^{r-1} \alpha_i \sigma_i^2 / r \text{ and } \bar{A}_t = \sum_{t}^{t+n-1} A_t / n.$$

If, the variations between σ_i's for different i's and between A_t's for different t's are not large, the ratio $\bar{\sigma}^2/\bar{A}_t$ will be approximately equal to var $(\epsilon_{it}) = \sigma_i^2/A_{t-i}$, which is expected to be small for large A_{t-i} or A_t since A_t's have not been assumed to differ widely with t. Thus when A_t is large var $\left(\frac{\sum\limits_{t}^{t+n-1} Q_{t+d}}{\sum\limits_{t}^{t+n-1} A_t} \right)$ is expected to be quite small at leastw hen $n \geqslant r$. Hence if the ratio $\sum\limits_{t}^{t+n-1} Q_{t+d} \Big/ \sum\limits_{t}^{t+n-1} A_t$ is worked out successively with higher values of n starting from $n = 2$ and it is noticed that for a certain n the variation of the above ratio becomes insignificant, i.e. the ratio assumes more or less a stable value with different t's (see Example 1), we may then conclude that $r \simeq n$. In such a case, therefore, we can arrive at indirectly an approximation to the number r of different batches of students present after a specific period of d years. This leads to an estimate of wastage in time taken by the students admitted into the training course.

With the proper choice of n, when the ratio $\sum\limits_{t}^{t+n-1} Q_{t+d} \Big/ \sum\limits_{t}^{t+n-1} A_t$ assumes a stable value, it may be possible to estimate $\tilde{p}$ from equation (2.9), provided a sufficient number of equations with varying t's (evidently this number will have to be greater than or equal to r) are available. Since the right hand side of equation (2.12) is independent of t, it is not possible to estimate separately $\tilde{p}_i$'s and we have to be satisfied with an estimate of $\sum\limits_{i=0}^{r-1} \alpha_i \tilde{p}_i$ only. In most of the practical situations (see Examples 1 and 2) α_i's are very nearly equal to 1, so that in such cases it is possible to estimate only the sum $\sum\limits_{i=0}^{r-1} \tilde{p}_i$ which leads to an estimate of the total wastage $1 - \sum\limits_{i=0}^{r-1} p_i$ during the entire training period.

It may also be mentioned that when A_{t-i} is not large which may be the case when the number of admissions in an individual college only are considered, the assumption that var (ϵ_{it}) is negligible may not be valid. In such cases the ratio $\sum_t^{t+n-1} Q_{t+d} \Big/ \sum_t^{t+n-1} A_t$ is expected to assume a stable value only when n is considerably greater than r; as such it is not possible here to get an estimate of r in an indirect way as is possible when $A_{t=i}$ is assumed to be large. Thus in such situations it is not possible to estimate the wastage in time but as in the previous case estimation of total wastage during the training period will usually be possible (see Example 2). It may also be pointed out here that the wastage in time for an individual training institution can be estimated without much difficulty from a detailed cohort study mentioned earlier. Fortunately for us, in order to get an estimate of the average wastage in time for all students admitted in any training institution of the country the simple method indicated above provides some estimate whereas just in such situations detailed cohort study for each training institution which could have given a more reliable estimate is not in general practically feasible.

Thus under the proposed method at first the ratio $\sum_t^{t+n-1} Q_{t+d} \Big/ \sum_t^{t+n-1} A_t$ has to be successively worked out upwards starting with $n = 1$ till the above ratio assumes a stable value. However, in practical situations when Q_{t+d} and A_t are both large numbers, computation of the above ratios for successive values of t and n becomes a little time-consuming. The question that will naturally come to one's mind is whether any simpler method can be suggested to eliminate successive summations of both the numerator and denominator of the ratio for various values of n. For this purpose, the following identity which is easy to verify will be useful.

$$\frac{\sum_1^n x_i}{\sum_1^n y_i} - \frac{1}{n}\sum_1^n \frac{x_i}{y_i} \equiv \frac{1}{n}\sum_{\substack{i,j=1\\ i<j}}^{n}\left(\frac{x_i}{y_i}-\frac{x_j}{y_j}\right)\left(\frac{y_i-y_j}{\sum_1^n y_k}\right) \qquad \ldots (2.14)$$

It is thus seen that L.H.S. of (2.14) vanishes either when y's are all equal or when the ratios x_i/y_i, $i = 1, \ldots, n$, are identical. Further, when the ratios x_i/y_i, $i = 1, \ldots, n$, do not differ widely among themselves and the same is true for y_i's, it is clear that the L.H.S. of (2.14) will be very small, especially when y_i's are of the same sign, say positive, since in that case $\left(y_i - y_j\right) \Big/ \sum_1^n y_k$ will be much less than 1, and the R.H.S. of (2.14) is expected to be very small.

Hence the ratio $\sum_t^{t+n-1} Q_{t+d} \Big/ \sum_t^{t+n-1} A_t$ is expected to be very nearly equal to the expression $\frac{1}{n}\sum_t^{t+n-1}\frac{Q_{t+d}}{A_t}$ if A_t's do not differ very much among themselves and so also are the ratios Q_{t+d}/A_t for different values of t. When A_t is very large or in the case of an individual institution it is very likely that A_t's will not be very much different for different t's. Also in practice, Q_{t+d}/A_t's will all be less than 1 and the variation among these ratios will not be much. As such in practical situations in the method suggested above, we may work with $\frac{1}{n}\sum_t^{t+n-1}\frac{Q_{t+d}}{A_t}$ in place of $\sum_t^{t+n-1} Q_{t+d} \Big/ \sum_t^{t+n-1} A_t$.

In practice, it might be found profitable to proceed as follows :

(1) At first the ratios Q_{t+d}/A_t are worked out for each t.

(2) Next the overall average of Q_{t+d}/A_t's is evaluated from the available data.

(3) Then the various averages of Q_{t+d}/A_t's with successive smaller values of n are worked out.

(4) As soon as the variation among averages based on $n-1$ values, say, is found to be significantly larger than the variation among the averages based on n values, further computation is stopped.

(5) It is concluded that the average of Q_{t+d}/A_t's based on n consecutive values will approximate the overall percentage of success and 1 minus this average provides an estimate of total wastage.

(6) Further the overall percentage of success is achieved in $d+n-1$ years instead of the prescribed period of d years, provided A_t's are not small numbers. This gives some estimate of the wastage in time during the training period.

The above steps will be clear from the examples furnished below :

3. Illustrations

Example 1: Total number of admissions of students in all the medical colleges in India during 1947-1950 and the number who qualified after 5 years, i.e. during 1952-1955 were as follows :

year (t)	number admitted (A_t)	number qualified after 5 years (Q_{t+d})	$\frac{Q_{t+d}}{A_t}$
(1)	(2)	(3)	(4)
1947	2625	2164	82.4
1948	2811	2299	81.8
1949	2609	2687	103.0
1950	2675	2359	88.2

The overall average is

$$\frac{1}{4}\sum_{1947}^{1950}\frac{Q_{t+5}}{A_t} = 88.8 \ (n = 4).$$

Also

$$\frac{1}{3}\sum_{1947}^{1949}\frac{Q_{t+5}}{A_t} = 89.1, \ \frac{1}{3}\sum_{1948}^{1950}\frac{Q_{t+5}}{A_t} = 91.0 \ (n = 3)$$

and

$$\frac{1}{2}\sum_{1947}^{1948}\frac{Q_{t+5}}{A_t} = 82.1, \ \frac{1}{2}\sum_{1948}^{1949}\frac{Q_{t+5}}{A_t} = 92.4, \ \frac{1}{2}\sum_{1949}^{1950}\frac{Q_{t+5}}{A_t} = 95.6 \ (n = 2).$$

Thus it is seen that the average values assume more or less a stable value for $n \geqslant 3$ and not for $n = 2$ and the stable value of the average is about 90 percent. Hence it is concluded that only about 90 percent, of the medical students admitted for the graduate course of about 5 years' duration eventually complete the prescribed courses in about $d+n-1 = 7$ years' time, so that the real wastage during training period is about 10 percent of the students admitted and the wastage in time for most of the students who qualify is at most 2 years.

Example 2 : In an Engineering College in West Bengal the following information about the number of admissions, number appearing after 4 years—the usual duration of training-and the number qualifying after 4 years are available:

year (t)	number of admissions (A_t)	number appearing after 4 years (Q'_{t+4})	$\frac{Q'_{t+4}}{A_t}$	number qualifying after 4 years (Q_{t+4})	$\frac{Q_{t+4}}{A_t}$
(1)	(2)	(3)	(4)	(5)	(6)
1948	162	153	94.4	119	73.5
1949	160	160	100.0	150	93.8
1950	166	152	91.6	139	83.8
1951	170	138	81.2	111	65.3
1952	165	142	86.0	121	73.3
1953	188	149	79.3	127	67.6
	1011	894		767	

The overall averages for columns (4) and (6) are

$$\frac{1}{6}\sum_{1948}^{1953} \frac{Q'_{t+4}}{A_t} = 88.8\,; \quad \frac{1}{6}\sum_{1948}^{1953} \frac{Q_{t+4}}{A_t} = 76.2 \ (n = 6).$$

The averages based on $n = 5$ are

$$\frac{1}{5}\sum_{1948}^{1952} \frac{Q'_{t+4}}{A_t} = 90.6, \quad \frac{1}{5}\sum_{1949}^{1953} \frac{Q'_{t+4}}{A_t} = 87.6$$

and

$$\frac{1}{5}\sum_{1948}^{1952} \frac{Q_{t+4}}{A_t} = 77.9, \quad \frac{1}{5}\sum_{1949}^{1953} \frac{Q_{t+4}}{A_t} = 76.8.$$

With smaller n the variation between the averages will considerably increase. These are not worked out, since as it has been pointed out already that for a single institution estimation of wastage in time (for which successive evaluation of the average for different n is required) under the assumed model might be quite misleading and it is advisable to estimate attrition through cohort study of the students admitted into the college. When data from a single institution are available, it is advisable to derive only the overall percentage, viz $\sum_t Q_{t+4}/\sum_t A_t$ directly without taking recourse to the alternative method of estimation, which involves more calculation. From the above data we may only conclude that about 90 percent of the students admitted sit for the final examination and about 76 percent of the students admitted eventually qualified as engineers from the above institution during the period 1948-57.

4. Discussion

It will be seen that the suggested method provides only an approximate idea of attrition during training period of students under certain realistic assumptions. However, sometimes it may be found (which is more likely to happen when a large number of years are considered) that a few years may be quite abnormal due to certain extraneous causes by showing a very much different percentage from those of other years. Under such situations the assumed model may not have practical justification and it may be advisable not to put too much reliance on the estimated measures of attrition in such situations. Estimates of wastage as may be derived by the suggested method should preferably be verified by means of detailed cohort studies of students from the time of their admissions till the time they all qualify in a few selected training institutions so chosen that they may be considered to provide a representative sample of all the institutions of the country.

Lastly, it may be mentioned that although in this paper study of attrition in technical education is only mentioned, the suggested method will be applicable in any form of education, technical or non-technical. But in case of technical education the problem of measurement of attrition is at present very important and as such only technical education has been stressed in the paper.

Paper received : *February, 1959.*

A NOTE ON THE EFFICIENCY OF DOUBLE SAMPLING FOR STRATIFICATION*

By M. V. JAMBUNATHAN
University of Mysore

SUMMARY. The formula for the efficiency of the double sampling design is given in this note in terms of the population parameters and the conditions under which this design is more efficient than simple random sampling have been stated in algebraic form.

1. Introduction

The theory of the method of double sampling for stratification was first given by Neyman (1938). According to this method, the large sample (or the first stage sample) serves as a basis for stratification of the population according to the value of an auxiliary character x, and at the second stage, stratified sampling is adopted for measuring the principal character y in a subsample, and then the population value of y is estimated. Neyman has also worked out (under certain simplifying conditions) the optimum allocation of the sampling units to the large and small samples. He does not, however, give the limiting conditions for this sampling design to be more efficient than simple random sampling, and just mentions that the decision to use the one or the other method should be taken only after knowing by numerical substitution, the variances under the two methods.

In this note, the formula for the efficiency of the double sampling design is given in terms of the population parameters, and the conditions under which this design is more efficient than simple random sampling have been stated in algebraic form. Taking the cost of a single measurement of x in the population to be unity, and the cost of a single measurement of y in the i-th stratum to be c_i^2, the formula for the efficiency E of the design relative to simple random sampling is expressible in the form

$$\frac{1}{E} = \frac{V_{(ds)}}{V_{(ran)}} = \left\{ \frac{\overline{\sigma}_c}{C\sigma} + \frac{\eta}{C} \right\}^2,$$

where σ^2 is the variance of y in the population, $C^2 = \Sigma\ P_i c_i^2$ is the average cost of a single measurement of y in the population, $\overline{\sigma}_c = \Sigma\ P_i c_i \sigma_i$ is the weighted sum of the strata standard deviations σ_i's, the respective weights being equal to the product of c_i and the corresponding stratum proportion P_i, and η is the correlation ratio of y on x. The use of the double sampling design with stratification will therefore result in gains if $\eta < C - \overline{\sigma}_c/\sigma$, and will result in loss when the reversed inequality holds. When the c_i's are all equal, the common value becomes C, while the weighted sum $\overline{\sigma}_c$ becomes equal to $C\overline{\sigma}$, where $\overline{\sigma}$ is the weighted average of the stratum standard deviations, the weights being the stratum proportions. The efficiency E is then given by the relation

$$\frac{1}{E} = \left\{ \frac{\overline{\sigma}}{\sigma} + \frac{\eta}{C} \right\}^2.$$

---

* Paper presented to the 44th Session of the Indian Science Congress, 1957.

2. The estimate and its variance under double sampling for stratification

The technique of double sampling, (or two-phase sampling as it is sometimes called) is used extensively in sample surveys as a measure of increasing the precision of the estimate. This method of double sampling is of great utility for estimating the population value of a character y, when an auxiliary character x correlated with y is available, which is easy and inexpensive to measure. A large preliminary sample is selected at random, wherein the auxiliary character x alone is measured. The information furnished by the sample is used to establish a relation between x and y in the population; then depending on this relation, a suitable estimating formula is employed for estimating the mean value of the principal character y in the population. This relation may be the ratio between x and y, or the regression of y on x, etc. The case where the preliminary sample is used to obtain the nature of the frequency distribution of x, and in particular, to stratify the population according to the value of the auxiliary character x, has been considered by Neyman (1938). In applying this method, we consider the range of variation of x as being divided into s intervals, namely, from x_0 to x_1, x_1 to x_2, ..., x_{s-1} to x_s. Let π_i be the part of the population π composed of the individuals for which $x_{i-1} \leqslant x < x_i$, for $i = 1, 2, \ldots, s$. This may be called the i-th stratum of the population. Let the proportion of the individuals of π belonging to i-th stratum be P_i.

The first sample of size n is the simple random sample of the population which is assumed large. From the measurements of x furnished by the elements of this sample, the strata proportions may be estimated. If n_i of the sampled individuals fall into the i-th stratum, then $p_i = n_i/n$ furnishes an unbiased estimate of P_i with a variance equal to P_iQ_i/n, where $Q_i = 1-P_i$. Also $\text{cov}(p_i, p_j) = -P_iP_j/n$. The second sample is a stratified random sample of size m (say), when measurements of y are taken. From among the n_i elements of stratum i, we draw at random a subsample of m_i elements; let the observed values of these elements be y_{ij} for $i = 1, 2, \ldots, s$; $j = 1, 2, \ldots, m_i$, so that $m = \Sigma m_i$. If the means of y for the several strata in the population be $\bar{Y}_i$, $i = 1, 2, \ldots, s$, and if the overall mean of y in the population be $\bar{Y}$, then

$$\bar{Y} = \Sigma \, P_i \bar{Y}_i. \qquad \ldots \quad (2.1)$$

An unbiased estimate of $\bar{Y}$ is given by

$$\bar{Y}' = \Sigma \, p_i \bar{y}_i, \qquad \ldots \quad (2.2)$$

where

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij},$$

the sample mean for the i-th stratum. The variance of this estimate is

$$V(\bar{Y}') = \sum_i \frac{\sigma_i^2}{m_i} \left(P_i^2 + \frac{P_iQ_i}{n} \right) + \frac{1}{n} \sigma_b^2, \qquad \ldots \quad (2.3)$$

where σ_i^2 is the variance of y in the i-th stratum and σ_b^2 is the variance between strata means.

3. Optimum allocation under double sampling

On the assumption that the cost of a single observation of y is the same for all strata, Neyman has obtained the optimum allocation for this sample design. But we may take a more general case and assume that the cost of a single measurement of y varies from stratum to stratum while that of a single measurement of x is constant. Then it can be shown (Jambunathan, 1953) that the optimum allocation of the sampling units to the large sample and to the different strata in the subsample is determined approximately by the equations

$$P_i\sigma_i/m_i\sqrt{a_i} = \sigma_b/n\sqrt{b}, \qquad \dots \quad (3.1)$$

leading to the explicit solutions

$$m_i = \frac{T_0}{S} \cdot \frac{P_i\sigma_i}{\sqrt{a_i}}\ ; \quad n = \frac{T_0}{S} \cdot \frac{\sigma_b}{\sqrt{b}}, \qquad \dots \quad (3.2)$$

where $S = (\Sigma P_i\sigma_i\sqrt{a_i})+\sigma_b\sqrt{b}$, and T_0 is the budgeted cost, excluding overheads.

The relations (3.1) and (3.2) given above are approximate, since, in deriving them, the terms containing $n\ m_i$ in the denominator have been neglected as being small. The minimum variance under the allocation (3.2) is then approximately given by

$$V_{\min} = \frac{S^2}{T_0} = \frac{[(\Sigma P_i\sigma_i\sqrt{a_i})+\sigma_b\sqrt{b}]^2}{T_0}. \qquad \dots \quad (3.3)$$

If however, the between and within strata variances are such as to make $\sigma_i^2/a_i > \sigma_b^2/b$, then the expected value of n_i for the i-th stratum is less than m_i. In that case, there is no subsampling in stratum i, and all the n_i individuals falling into the i-th stratum during the first stage of sampling are included into the second sample; as a consequence, optimum allocation to the remaining strata should be re-worked as in the case of ordinary stratified random sampling (Jambunathan, 1953, p. 38). A fuller discussion of this special case is contained in a paper presented by the author to the 45th Session of the Indian Science Congress, Madras (Jambunathan, 1958).

4. Relative efficiency of the double sampling design

In order to understand the nature of the function for $V_{\min}$, and to compare its value with that of the variance of the estimate under simple random sampling, let us put $a_i = bc_i^2$; that is to say, the unit cost of measuring y in the i-th stratum is c_i^2 times the unit cost of measuring x. The minimum variance can then be written as

$$V_{\min} = b(\Sigma\ P_ic_i\sigma_i+\sigma_b)^2/T_0 \qquad \dots \quad (4.1)$$

$$= b(\bar{\sigma}_c+\eta\sigma)^2/T_0, \qquad \dots \quad (4.2)$$

where η is the correlation ratio of y on x, and $\bar{\sigma}_c = \Sigma P_ic_i\sigma$. Thus the efficiency of the design is dependent upon the correlation ratio and the method has the advantage of being more general, since no assumption regarding the nature of the regression of y on x is made. We shall now obtain an expression for the efficiency of this design relative to simple random sampling.

When a simple random sample is employed to estimate $\bar{Y}$ the expected value of the cost of including a single element into the sample would be $a = \Sigma P_i a_i$. Therefore, on the average, T_0/a elements could be included into the random sample for the same budgeted cost T_0, and the value of the variance of the estimate of $\bar{Y}$ obtained from such a random sample would be given by

$$V_{\text{ran}} = a\sigma^2/T_0 = b\, C^2\sigma^2/T_0, \qquad \dots \quad (4.3)$$

where C^2 denotes the ratio of a to b. If the relative efficiency of the double sampling design compared to simple random sampling be denoted by E, we have

$$\frac{1}{E} = \frac{V_{\min}}{V_{\text{ran}}} = \left\{ \frac{\bar{\sigma}_c}{C\sigma} + \frac{\eta}{C} \right\}^2. \qquad \dots \quad (4.4)$$

Therefore, double sampling with stratification will, on the average, result in gains if $1/E < 1$, that is, if $\eta < C - \bar{\sigma}_c/\sigma$, and will result in loss if otherwise. Taking all the c_i's to be equal to each other, the common value therefore becoming equal to C, the expression for the reciprocal of the efficiency becomes

$$\frac{1}{E} = \left\{ \frac{\bar{\sigma}}{\sigma} + \frac{\eta}{C} \right\}^2, \qquad \dots \quad (4.5)$$

where $\bar{\sigma}$ is the weighted average of the stratum standard deviations, the weights being proportional to the stratum sizes. In this case, the above double sampling design will lead to gains if

$$\eta < C(1-\bar{\sigma}/\sigma) \quad \text{or if} \quad C > \eta/(1-\bar{\sigma}/\sigma),$$

and will lead to loss if

$$C < \eta/(1-\bar{\sigma}/\sigma).$$

These results indicate that the efficiency of the design hinges upon the correlation ratio of y on x, and the weighted average of the stratum standard deviations relative to the standard deviation of the population as also on the average cost of measuring y relative to x.

The author wishes to express his thanks to the referee of the Indian Statistical Institute for some of the suggestions for revising the paper.

References

Neyman, J. (1938): Contribution to the theory of sampling human populations. *J. Amer. Stat. Ass.*, **33**, 101–116.

Jambunathan, M. V. (1953): *Some Aspects of Sampling*, India Book Company, Bangalore City.

——— (1958) : Optimum allocation with double sampling for stratification. *Proc. 45th Indian Science Congress*, Part III, Section 2 : Statistics.

Paper received : September, 1957.

Revised : August, 1958.

ON THE STANDARD ERROR OF THE LORENZ CONCENTRATION RATIO

By N. SREENIVASA IYENGAR
Indian Statistical Institute, Calcutta

SUMMARY. In this paper a large sample expression is derived for the standard error of the Lorenz concentration ratio on the assumption of log-normality and an empirical investigation is made to test the equality of two Lorenz ratios obtained from two interpenetrating sub-samples of the National Sample Survey.

1. INTRODUCTION

The object of this paper is to provide, on the assumption of log-normality, a large sample test for the Lorenz criterion by showing that its sampling distribution is asymptotically normal with its mean at its true value and variance involving the standard deviation of the log-normal distribution. A test of significance is also provided for the difference between concentration ratios and is numerically illustrated with respect to data on consumer expenditure in India.

2. THE LORENZ RATIO

In order to avoid the graphical calculation of the Lorenz ratio various theoretical approximations have been attempted from time to time since the Pareto's law of incomes. Aitchison and Brown (1957) have recently considered log-normal description for the distribution of incomes along side Pareto's and in the former case they have given expression for L, the Lorenz ratio, in terms of the standard deviation of the log-normal distribution. For purposes of completeness, however, we shall define L for the concentration of the consumer expenditure assuming that the per capita consumption expenditure (x) obeys a log-normal distribution with mean θ and standard deviation λ, i.e.

$$f(x;\theta,\lambda)\;dx = \frac{1}{x\lambda\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{\log x-\theta}{\lambda}\right)^2}\,dx,\;\; 0<x<\infty \qquad \ldots \quad (2.1)$$

and having the mean, $e^{\theta+\frac{1}{2}\lambda^2}$, and variance, $e^{2\theta+\lambda^2}(e^{\lambda^2}-1)$.

Let $p(x)$ be the proportion of persons having a per capita expenditure not more than x each and $q(x)$, the proportion of total expenditure of these persons. We have

$$p(x) = \int_0^x f(x;\;\theta,\;\lambda)dx$$

$$q(x) = \int_0^x xf(x;\theta,\lambda)dx \div \int_0^\infty xf(x;\theta,\lambda)dx$$

which simplify respectively to

$$p(x) = \varphi\left(\frac{\log x-\theta}{\lambda}\right),\; q(x) = \varphi\left(\frac{\log x-\theta}{\lambda}-\lambda\right) \qquad \ldots \quad (2.2)$$

where φ is defined by

$$\varphi(\xi) = \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt. \qquad \ldots \quad (2.3)$$

Equations in (2.2) represent the parametric equations to the concentration curve in which p is measured along the horizontal axis and q along the vertical axis. Elimination of x between them often leads to explicit equation of the concentration curve of the form $q = h(p)$. The Egalitarian line or the line of equal distribution is defined as the diagonal, $p = q$, on the concentration diagram. (See Figure 1 on p. 376).

The area between the Egalitarian line and the Lorenz curve is taken as a measure of concentration. The concentration ratio is defined as twice this area in order that it may take values between zero and unity. Since $p > q$ the Lorenz curve lies below the line of equal distribution and, therefore, the concentration ratio L becomes

$$L = 1-2A, \qquad \ldots \quad (2.4)$$

where

$$A = \int_0^1 q \, dp. \qquad \ldots \quad (2.5)$$

The integral in (2.5) can easily be worked out by putting $\log x - \theta = \lambda t$ and using the relations in (2.2). That is

$$A = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t-\lambda} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \right) \frac{e^{-\frac{1}{2}t^2}}{\sqrt{2\pi}} dt \qquad \ldots \quad (2.6)$$

which is simply the probability that the difference $x_1 - x_2$ does not exceed $-\lambda$ where x_1 and x_2 are any two independent observations from the unit normal distribution. Now since $x_1 - x_2$ has a mean zero and standard deviation $\sqrt{2}$ the probability is simply $\varphi\left(-\frac{\lambda}{\sqrt{2}}\right)$ in the notation of (2.3). Hence we have

$$L = 1-2\varphi\left(-\frac{\lambda}{\sqrt{2}}\right) = 2\varphi\left(+\frac{\lambda}{\sqrt{2}}\right)-1 \qquad \ldots \quad (2.7)$$

which is a function of λ alone.

3. The estimate of L and its standard error

It is easily seen that L is a continuous function of λ and its derivatives are also continuous. Under these circumstances the maximum likelihood estimate of L is simply provided by replacing λ in equation (2.7) by its maximum likelihood estimate $\hat{\lambda}$. Moreover, $\hat{L}$, the maximum likelihood estimate of L, has an asymptotic normal distribution with mean L and variance given by

$$V(\hat{L}) = \frac{\lambda^2 e^{-\frac{1}{2}\lambda^2}}{2\pi n}. \qquad \ldots \quad (3.1)$$

Taking positive square roots on both sides we have

$$\text{s.e.}\ (\hat{L}) = \frac{\hat{\lambda} e^{-\frac{1}{4}\hat{\lambda}^2}}{\sqrt{2\pi n}}. \qquad \ldots \quad (3.2)$$

This is calculated by first computing $\hat{\lambda}/\sqrt{n}$ and multiplying it by the ordinate corresponding to $\hat{\lambda}/\sqrt{2}$ of the unit normal curve.

For a grouped frequency data it is rather tedious to apply the maximum likelihood method of estimation since information on individuals generally is not available. In such cases one may follow any one of a number of simpler methods of estimating λ based on the empirical Lorenz curve. For example, it is easy to see that at the mean per capita expenditure, $p = \varphi\left(\frac{\lambda}{2}\right)$ and $q = \varphi\left(-\frac{\lambda}{2}\right)$ so that $p+q = 1$. The latter is the other diagonal of the Lorenz diagram and the abscissa corresponding to the point where it intersects the Lorenz curve gives the proportion of individuals who are below the average per capita expenditure. If this proportion be p_0 then the value of λ is simply $2\eta_0$ where η_0 satisfies the relation $p = \varphi(\eta)$. It is not known whether this method is a good approximation to the maximum likelihood method from ungrouped data. The expression (3.2), therefore gives an underestimate of the standard error. If, however, one follows the maximum likelihood method for grouped data as is the practice in the probit analysis the above expression will be a very close lower bound to the true expression for standard error; there will still be some difference due to loss of efficiency by grouping.

4. A test for equality of concentration

We have seen in the previous section that Lorenz concentration ratio and its sampling error can be measured in terms of the variance of the log-normal distribution but independent of the location parameter. Therefore, any statistical test for the significance of the concentration ratio is equivalent to the test for the significance of the standard deviation computed from the logarithms of the expenditures, having a normal distribution. More generally, if there are two or more sources from which the concentration ratios are estimated, a test can be provided for their equality and hence for consistency of the different sources. From the concentration curves for each source the variance parameter, namely, λ^2 is estimated as in section 3 and Bartlett's test is employed for the equality of several variances. In the case of two sources, as for example, two sub-samples in any survey, the problem is quite simple. For in this case an exact test can be performed on the ratio $\hat{\lambda}_1^2/\hat{\lambda}_2^2$ by entering in the F-Table with appropriate degrees of freedom. However, in large samples we can use a normal test for the difference, $(\log \hat{\lambda}_1 - \log \hat{\lambda}_2)$.

5. A NUMERICAL ILLUSTRATION

For purposes of illustration we consider the following table chosen from Mahalanobis's (1958) "Science and National Planning." The two independent sources refer to the two independent sub-samples of the National Sample Survey.

TABLE 1. CUMULATIVE PERCENTAGE OF EXPENDITURE IN RUPEES OF "ALL" ITEMS : ALL INDIA (RURAL AND URBAN) 1955-56

per capita expenditure class Rs. (x)	cumulative percentage of total number of persons (p)		cumulative percentage of expenditure on 'all' items (q)		average expenditure in Rs.
	s.s. 1	s.s. 2	s.s. 1	s.s. 2	
(1)	(2)	(3)	(4)	(5)	(6)
1. 0—8	13.84	14.18	4.67	4.73	6.26
2. 8—11	31.52	31.12	13.55	13.26	9.41
3. 11—13	43.62	42.50	21.38	20.45	11.98
4. 13—15	56.65	51.74	28.23	27.71	13.96
5. 15—18	64.17	63.33	38.57	37.20	16.49
6. 18—21	73.61	70.32	48.59	44.36	19.54
7. 21—24	79.36	78.63	55.62	54.19	22.51
8. 24—28	85.93	83.14	64.78	60.33	25.79
9. 28—34	90.28	89.12	71.88	70.10	30.68
10. 34—43	95.31	93.79	82.17	79.36	37.72
11. 43—55	97.08	97.19	86.85	87.69	47.24
12. 55 & over	100.00	100.00	100.00	100.00	100.00
all classes	—	—			18.74

Concentration curves are drawn for both sub-samples taking cumulative percentage of total number of persons on the horizontal axis and the cumulative percentage of expenditure on "all" items on the vertical axis. The curves are sufficiently close to each other and one would naturally feel tempted to conclude that the two sub-samples are not inconsistent as shown by the area between the thick line and the dotted line (Figure 1).

For estimation of λ two perpendiculars are dropped from where the two curves intersect the diagonal opposite the Egalitarian line. For the sub-samples 1 and 2, the corresponding abscissae are found to be $p_1 = 0.6212$ and $p_2 = 0.6216$, which correspond to the mean expenditures in the two sub-samples.

The λ's are immediately calculated from the relation, $\lambda = 2\varphi^{-1}(p)$ where $\varphi^{-1}(p)$ is the standard normal deviate corresponding to p. In the present example we have $\hat{\lambda}_1 = 0.6133$, and $\hat{\lambda}_2 = 0.6194$.

Before proceeding to the estimation of the Lorenz ratio and its standard error we shall first examine the adequacy of the log-normal hypothesis. This is done by plotting the η-probit of p defined by

$$p(x) = \varphi(\eta) \qquad \ldots \quad (5.1)$$

against $\xi = \log_e x$. For both sub-samples 1 and 2 resulting graphs are almost straight lines. The necessary computations are, however, presented in the following scheme.

TABLE 2. CUMULATIVE PERCENTAGE OF PERSONS

per capita expenditure class Rs. (x)	less than Rs. (x)	cumulative proportion of persons		$\log_e x$	η-probit	
		s.s. 1	s.s. 2		s.s. 1	s.s. 2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0— 8	8	0.1384	0.1418	2.0794	−1.0893	−1.0714
8—11	11	0.3152	0.3112	2.3979	−0.4817	−0.4930
11—13	13	0.4362	0.4250	2.5649	−0.1611	−0.1891
13—15	15	0.5665	0.5174	2.7080	0.1662	0.0426
15—18	18	0.6417	0.6333	2.8904	0.3638	0.3398
18—21	21	0.7361	0.7032	3.0445	0.6311	0.5330
21—24	24	0.7936	0.7863	3.1780	0.8204	0.7926
24—28	28	0.8593	0.8314	3.3322	1.0758	0.9581
28—34	34	0.9028	0.8912	3.5264	1.2988	1.2319
34—43	43	0.9531	0.9379	3.7612	1.6747	1.5382
43—55	55	0.9708	0.9719	4.0073	1.8957	1.9110
55 & over	∞	1.0000	1.0000			

The graphs are drawn for both sub-samples 1 and 2 using the dotted line for sub-sample 1 and thick line for sub-sample 2, and taking ξ on the horizontal axis and η on the vertical axis.

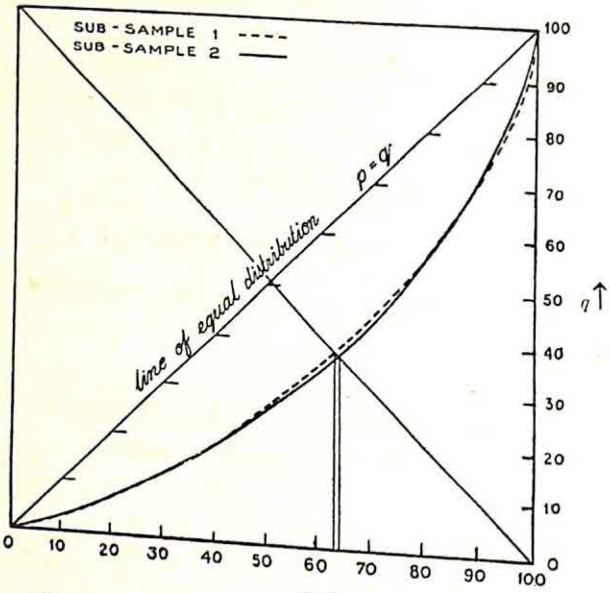


Figure 1. Lorenz curves of total expenditure

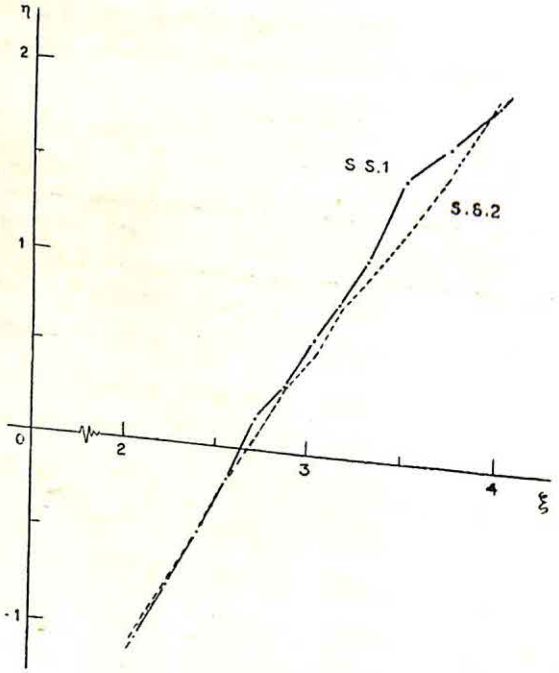


Figure 2. Distribution of per capita expenditure: Log-probit Scale

It will be seen from Figure 2 that the graphs, besides being approximately linear, are not too divergent, suggesting thereby that the two sub-samples are probably consistent. This point will be examined later. However, Table 3 is given below to show the adequacy of log-normal fit in both cases.

TABLE 3. THEORETICAL PERCENTAGE DISTRIBUTION OF LOG-PER CAPITA EXPENDITURE

log-per capita expenditure	percentage			
	s.s. 1		s.s. 2	
	theoretical	actual	theoretical	actual
1. $-\infty$ to 2.0794	13.98	13.84	14.35	14.18
2. 2.0794 — 2.3979	14.73	17.68	14.75	16.94
3. 2.3979 — 2.5649	9.89	12.10	9.85	11.38
4. 2.5649 — 2.7080	9.15	13.03	9.17	9.24
5. 2.7080 — 2.8904	11.77	7.52	11.65	11.59
6. 2.8904 — 3.0445	9.35	9.44	9.25	6.99
7. 3.0445 — 3.1780	7.24	5.75	7.17	8.31
8. 3.1780 — 3.3322	7.07	6.57	7.01	4.51
9. 3.3322 — 3.5264	6.76	4.35	6.72	5.98
10. 3.5264 — 3.7612	5.22	5.03	5.24	4.67
11. 3.7612 — 4.0073	2.89	1.77	2.91	3.40
12. 4.0073 and above	1.95	2.92	2.03	2.81
	100.00	100.00	100.00	100.00

Having now examined the assumption of log-normality we next estimate the Lorenz concentration ratios, L_i $(i = 1, 2)$ using the relation, $L_i = 2\varphi\left(\frac{\lambda_i}{\sqrt{2}}\right) - 1$. Substituting the values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ we have at once $\hat{L}_1 = 0.3355$ and $\hat{L}_2 = 0.3386$ so that the area between the two concentration curves is 0.0016, being half of the difference between $\hat{L}_1$ and $\hat{L}_2$. Finally their standard errors are found to be 0.00317 and 0.00325 respectively.

Since the sample size (7023) of individuals is large it is enough to employ the normal test for examining the difference in the variances (and hence concentration ratios). We now compute the positive difference, $\sqrt{7023}\,(\log \hat{\lambda}_1 - \log \hat{\lambda}_2)$ which simplifies to 0.5865. The upper $2\frac{1}{2}\%$ of the unit normal curve is about 1.96 and, therefore, it will be concluded that the difference in concentration between the two sub-samples is not significant on the 5% level. In other words, we do not have evidence to doubt the consistency of the two sub-samples in as far as they provide a valid estimate of the concentration ratio.

The NSS (National Sample Survey) practice is to select households through a rather complicated design, involving stratification, multistage sampling and selection with varying probabilities. The use of 7023 as the sample size is therefore open to serious objections. It seems, however, that the conclusion in the present case stands on very strong grounds. If one assumes, as a rough approximation, that the NSS samples behave like a random sample of size, say n, for purposes of estimating the distribution of persons over the per capita expenditure groups, then obviously this n, the effective sample size cannot seriously exceed 7023, as a matter of fact it is believed to be very much lower. In any case n cannot exceed 82,000 $[= 7023\ (1.96 \div 0.5865)^2]$ so that the criterion, $\sqrt{n}\,(\log \hat{\lambda}_1 - \log \hat{\lambda}_2)$ is less than 1.96. In other words, the difference is not statistically significant.*

6. Conclusion

The foregoing results should be of importance in the study of consumption patterns from time to time, provided it is empirically shown that the underlying distributions obey log-normality. There, are, however, some other theoretical considerations which should not be ignored in practical applications. Thus, the results could be misleading if the data were not fully representative. It is actually necessary that each individual be chosen at random with equal probability. This is perhaps far from true for the present data. Another possible source of error is the approximation to the method of maximum likelihood for ungrouped data by the method actually employed for the grouped data.

I am greatly indebted to Shri M. Mukherjee and Dr. D. Basu for their valuable suggestions and help in preparing this paper.

References

Aitchison, J. and Brown, J. A. C. (1957): *The Log-normal Distribution with Special Reference to Its Uses in Economics*, Cambridge University Press.

Mahalanobis, P. C. (1958): Science and National Planning. President's Anniversary Address, National Institute of Sciences of India, reprinted in *Sankhyā*, **20**, 69–106.

Paper received : August, 1958.

---

* I am grateful to Shri N. Bhattacharya for these remarks.

ON A PROBLEM OF ESTIMATING INCREASE IN CONSUMER DEMAND*

By N. SREENIVASA IYENGAR

Indian Statistical Institute, Calcutta

SUMMARY. This paper studies the effect of changes in concentration of income of individuals on the projection of future levels of consumption of both necessaries and luxuries. It is shown that, for example, if one does not take into account the changes in the income distribution in the direction of Egalitarianism one is likely to underestimate the need for necessary goods and services. This and a few other results are derived under the assumptions of log-normality of the income distribution and constancy of the Engel elasticity for all incomes. These results are numerically illustrated from the Indian National Sample Survey data.

1. Introduction

Under a planned programme of governmental investment the real income of individuals is normally expected to rise. The rise in real income will, in turn, lead to increase in demand for consumer goods and services. Unless the productive forces in the country are mobilised to meet the increase in demand, even partially, prices may sour up and ultimately result in serious inflation—a situation which is undesirable. For this and for other reasons estimates of expected increase in consumption of different items are quite useful in planning.

If the pattern of expenditure by individuals at different levels of income is known and the assumption is made that when an individual moves on to a higher income group, his pattern of expenditure would tend to be the same on an average as that of an individual who already belongs to that higher income group, it is possible to estimate the expected increase in demand for any given change in the distribution of income.

In India, however, complete data on incomes are not available and consequently we are tempted to use data on overall consumer expenditure as proxy for income. This is to a considerable extent justified since in India the average per capita saving is negligible. There exists a wealth of data on consumer expenditure since the emergence of the National Sample Survey (NSS) in 1950-51.

A general formulation of this problem was given by Roy and Laha (1959) who assumed constant elasticity and Pareto's form for expenditure distribution in estimating the relative increase in demand of certain goods, using material collected from the fourth round of the NSS. Later Roy and Dhar (1959) tried, in addition to constant elasticity curves, several of Tornquist's forms for Engel curves and the log-normal hypothesis for the distribution and estimated the relevant parameters from the seventh round materials of the National Sample Survey. They derived formulae for estimating increase in demand due to infinitesimal changes in the parameters, but did not derive any actual estimates. A graphical method of computing Engel elasticities from concentration curves was derived by Iyengar (1960) under the assumption of log-normality of the distribution of expenditure.

---

*A part of this paper was read at the Preliminary Conference of the Indian Members of the Econometric Society held at the Indian Statistical Institute, Calcutta, on 28 January 1960.

This paper assumes log-normality of the expenditure distribution and constant elasticity of Engel curves and presents a formula for obtaining numerical estimates of the expected increase in demand for five different items of consumer expenditure under a particular set of specifications about the changes desired in the expenditure distribution resulting from governmental investment plans. The estimates of relevant parameters are obtained from the National Sample Survey, tenth round (December 1955–May 1956).

2. Formulation of the general problem

If there exists a set of known relationships between overall expenditure and average expenditure per person on different items of the family budget then it may be easy to calculate the expected increase in expenditure on any particular item provided one can assume that, for any postulated distribution of the overall expenditure, such a relationship is invariant and the expenditure distribution remains unaltered in its form although undergoing some changes in its essential parameters.

Denoting by x the overall expenditure per capita and by y the expenditure per capita on a particular item let us write the marginal distribution of x by $g(x)$ and the conditional expectation of y for a given value of x by $E(y|x) = \psi(x)$. Now suppose that the plan investment alters the marginal distribution of x from $g(x)$ to $g^*(x)$. Then under the assumption that the function ψ remains invariant, the percentage increase in the average expenditure per capita on the given item is given by

$$I = 100 \left[\frac{E(\psi \,|\, g^*)}{E(\psi \,|\, g)} - 1 \right] \qquad \ldots \quad (2.1)$$

where

$$E(\psi \,|\, g) = \int_0^\infty \psi(x) g(x) dx \qquad \ldots \quad (2.2)$$

is the average expenditure per capita on the particular item when the overall expenditure distribution is $g(x)$ and similarly for $E(\psi \,|\, g^*)$.

We shall set down that the autonomous plan investment will lead to $100\,\alpha$ percent increase in the overall expenditure per capita of the community while at the same time the inequality of the overall expenditure distribution diminishes by $100\,\beta$ percent, the inequality being measured by the Lorenz ratio. Under these specifications it is possible to calculate the new parameters in g^* in terms of the parameters of g and the given proportions α and β.

3. The prediction formula

Suppose we adopt a constant elasticity hypothesis regarding the modified Engel curve. That is, we assume,

$$\psi(x) = Ax^{\eta} \qquad \ldots \quad (3.1)$$

where η is the expenditure elasticity of the specific item under consideration. Let us further suppose that the expenditure distribution can be adequately represented by a log-normal distribution with parameters θ and λ (say) and it changes in such a manner that it remains log-normal but with different parameters θ^* and λ^*. In other words,

$$g(x) = \begin{cases} \dfrac{1}{x\lambda\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\log x-\theta}{\lambda}\right)^2} & \text{if } x > 0 \\ 0 & \text{if } x \leqslant 0\,. \end{cases} \qquad .. \quad (3.2)$$

This distribution has the mean, $E(x) = e^{\theta+\frac{1}{2}\lambda^2}$, and the Lorenz concentration ratio, $L = 2\phi\left(\frac{\lambda}{\sqrt{2}}\right)-1$, (Aitchison and Brown, 1957, p. 112), where ϕ is given by

$$\phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2}\, du. \qquad \ldots \quad (3.3)$$

Under these assumptions the expression (2.1) of the last section simplifies to

$$I = 100 \left[\, e^{(\theta^*-\theta)\eta+\frac{1}{2}(\lambda^{*2}-\lambda^2)\eta^2} - 1 \,\right]. \qquad \ldots \quad (3.4)$$

This result can be established on observing that for log-normal variable the powers and multiples are also log-normal (Aitchison and Brown, 1957, p. 11) that is, $\psi(x) = Ax^{\eta}$ is also log-normal with the mean

$$E(\psi\,|\,g) = A\, e^{\eta\theta+\frac{1}{2}\eta^2\lambda^2}. \qquad \ldots \quad (3.5)$$

The expression in (3.4) involves two unknown parameters θ^* and λ^*. However, under the specifications in the last paragraph of Section 2 we have the following two relations giving θ^* and λ^* in terms of the original parameters and the proposed values of α and β.

$$e^{\theta^*+\frac{1}{2}\lambda^{*2}} = (1+\alpha)\, e^{\theta+\frac{1}{2}\lambda^2} \qquad \ldots \quad (3.6)$$

$$L^* = (1-\beta)L. \qquad \ldots \quad (3.7)$$

The new parameter λ^* can be readily computed from the relation,

$$\lambda^* = \sqrt{2}\, t_{\frac{1}{2}(1+L^*)} \quad \dots \quad (3.8)$$

since L^* is known as soon as β and L are known. Here t is given by the equation,

$$\phi(t_k) = k, \quad 0 \leqslant k \leqslant 1 \quad \dots \quad (3.9)$$

where ϕ is defined in (3.3) above. Using the result of (3.6) we can simplify (3.4);

$$I_{\alpha,\beta}\ (\eta) = 100\, [C(\beta, \eta)\, (1+\alpha)^{\eta} - 1] \quad \dots \quad (3.10)$$

where

$$C(\beta, \eta) = e^{\frac{1}{2}\eta(1-\eta)\, (\lambda^2 - \lambda^{*2})} \quad \dots \quad (3.11)$$

is ultimately a function of β and η since λ^* involves β when the initial concentration is known.

As a particular case if we assume that the expenditure distribution has not changed apart from a proportionate increase in the per capita total expenditure, i.e., if $\beta = 0$, it follows from (3.11) that the C-coefficients reduce to unity ierrspective of whether the commodity is a luxury item or a necessary item. In this case formula (3.10) becomes

$$I_{\alpha,0}\ (\eta) = 100[(1+\alpha)^{\eta} - 1], \quad \dots \quad (3.12)$$

a form which does not require that the distribution of total per capita expenditure be log-normal (Roy and Laha, 1959). It is easy to verify the following inequalities :

$$I_{\alpha,0}\ (\eta) < I_{\alpha,\beta}\ (\eta) \quad \text{if} \quad 0 < \eta < 1$$

$$I_{\alpha,0}\ (\eta) > I_{\alpha,\beta}\ (\eta) \quad \text{if} \quad \eta > 1. \quad \dots \quad (3.13)$$

These inequalities as well as those which appear subsequently are easily proved by observing that by definition $C(\beta, \eta)$, for $\beta > 0$, exceeds unity if $0 < \eta < 1$ while it falls short of unity if $\eta > 1$ (see Table 2).

On the other hand if one were to assume that the average income (total expenditure) is more or less stable through time and that the income distribution has slightly changed towards Egalitarianism, one gets a positive increase in demand for necessary goods and a fall in demand for luxuries. In symbols,

$$I_{0,\beta}\ (\eta) > 0 \quad \text{if} \quad \beta > 0, \quad 0 < \eta < 1$$

$$I_{0,\beta}\ (\eta) < 0 \quad \text{if} \quad \beta > 0, \quad \eta > 1. \quad \dots \quad (3.14)$$

It is also interesting to note that

$$I_{0,\beta}\ (\eta) < I_{\alpha,\beta}\ (\eta) \quad \dots \quad (3.15)$$

irrespective of the nature of the commodity. This result holds even if the income distribution has tended to be more concentrated. The decrease in demand for luxuries varies with the degree to which they can be considered as luxuries. In other words, if s and s' are two items of luxury, i.e.,

$$\eta_s, \eta_{s'} > 1$$

then

$$I_{0,\beta}(\eta_{s'}) < I_{0,\beta}(\eta_s) \quad \text{if } \eta_{s'} > \eta_s > 1. \qquad \ldots \quad (3.16)$$

In the same manner we may obtain, for all necessary goods s and s' for which $0 < \eta_s, \eta_{s'} < 1$ the following :

$$I_{0,\beta}(\eta_s) > I_{0,\beta}(\eta_{s'}) \quad \text{if} \quad \eta_s < \eta_{s'}. \qquad \ldots \quad (3.17)$$

If, however, the inequality of income distribution has increased as a result of planning (This is a fact not generally welcomed by planners but nevertheless the possibility of such an occurrence cannot be ruled out), i.e., $\beta < 0$, the above conclusions will be reversed in most of the above inequalities. In that case we will have the following general inequalities :

$$I_{\alpha,-\beta}(\eta) < I_{\alpha,0}(\eta) < I_{\alpha,\beta}(\eta) \quad \text{for all } 0 < \eta < 1$$

$$I_{\alpha,-\beta}(\eta) > I_{\alpha,0}(\eta) > I_{\alpha,\beta}(\eta) \quad \text{for all } \eta > 1. \qquad \ldots \quad (3.18)$$

If the changes in the income distribution were not taken into account and the assumption be made that the incomes of persons would increase by the same proportion and that the income elasticity is constant throughout the period of increase we obtain a different formula which is used by some practical workers. We shall distinguish this estimate and write I^* where I^* is given by

$$I^*_{\alpha,0}(\eta) = 100\,\alpha\eta. \qquad \ldots \quad (3.19)$$

Assuming that α is positive it is easy to obtain the following results* :

$$I_{\alpha,0}(\eta) > I^*_{\alpha,0}(\eta) \quad \text{if } \eta > 1$$
$$I_{\alpha,0}(\eta) < I^*_{\alpha,0}(\eta) \quad \text{if } 0 < \eta < 1 \qquad \ldots \quad (3.20)$$

the equality, however, holding when $\eta = 1$.

It will be seen below from Figure 1 presented in Section 4 by way of illustration from the National Sample Survey, that although the estimates $I^*_{\alpha,0}$ deviate from $I_{\alpha,0}$ in the correct direction, i.e., towards $I_{\alpha,\beta}$, where β is positive, these $I_{\alpha,0}$ values cannot be taken as substitutes for $I_{\alpha,\beta}$.

---

*These inequalities follow from a well known inequality :

For

$$x > y > 0.$$

$$mx^{m-1}(x-y) \underset{<}{\geqslant} x^m - y^m \underset{<}{\geqslant} my^{m-1}(x-y)$$

according as $0 > m > 1$, $m = 1$ or 0, or $0 < m < 1$. In this if we put $x = 1+\alpha$, $y = 1$ and $m = \eta$, we obtain the desired result (3.20). This point was suggested by Shri R. Narayanan.

4. Some illustrative examples from National Sample Survey data

We shall choose for purposes of illustration the following five important items of consumer expenditure which include necessaries as well as relative luxuries: (a) foodgrains, (b) all food, (c) clothing, (d) milk and milk products and (e) medicine and medical services. Their expenditure elasticities and specific concentration ratios are given in the following table.

TABLE 1. MODIFIED ENGEL ELASTICITIES AND SPECIFIC CONCENTRATION RATIOS

item (s)	Engel elasticity (η_s)	specific concentration ratio (L_s)
(1)	(2)	(3)
(a) foodgrains	0.58 (necessary)	0.21
(b) all food	0.76 ,,	0.28
(c) clothing	1.55 (luxury)	0.53
(d) milk and milk products	1.67 ,,	0.57
(e) medicine and medical services	1.95 ,,	0.64

These ratios and elasticities are computed from the National Sample Survey data on consumption collected in the tenth round covering the rural and urban areas of India and roughly correspond to the year 1955-56. The method of computation of the Engel elasticities and of the concentration ratios is given in detail in an earlier paper (Iyengar, 1960). These figures are to be taken only as illustrative and not as the NSS estimates. Our computations involved the use of a concentration map consisting of specific concentration curves and the Lorenz curve which together describe the average consumption pattern of the community. Lorenz curve is one which relates the proportions of total expenditure to the proportions of persons spending up to a given level of total expenditure per capita while the specific concentration curve is defined as showing the proportions of total consumption (value) of the specific commodity against proportions of persons spending up to a given level of total expenditure per capita. The specific concentration ratios are defined, in analogy with the Lorenz ratio, as twice the areas bound by the concentration curves and the Egalitarian line and these ratios may be used in characterizing commodities according as whether they are necessaries or luxuries. Under the assumptions stated in Section 3, the specific concentration ratio L_s is given by $2\phi\left(\frac{\lambda\eta_s}{\sqrt{2}}\right)-1$ where η_s is the Engel elasticity of the specific commodity s. The Lorenz ratio L is, of course, defined earlier and equals $2\phi\left(\frac{\lambda}{\sqrt{2}}\right)-1$. The item s is a necessary if $L_s < L$, and a luxury if $L_s > L$; in the former case, the specific concentration curve lies entirely above the Lorenz curve, and below in the latter case (Roy, Chakravarti and Laha, 1959). For instance, the first two items of Table 1 above are necessaries for which the concentration ratios are less than 0.36 and the last three items are luxuries since their concentration ratios exceed 0.36.

The above ratios may be empirically computed independent of the assumptions regarding the distributions. Thus, if p_i be the proportion of population whose per capita total expenditure does not exceed x_i, the latter being the upper value of the i-th total per capita expenditure class, q_i and Q_i be the corresponding proportions of total expenditure and specific commodity consumption respectively by the above stratum of population, ($i = 1, 2, \ldots, m$, say) then the specific concentration ratio is approximately given by

$$L_s = 1 - \sum_{i=1}^{m} (p_i - p_{i-1})(Q_i + Q_{i-1}) \qquad \ldots \quad (4.1)$$

where $(p_0, Q_0) = (0, 0)$ and $(p_m, Q_m) = (1,1)$. The Lorenz ratio L is computed in the same manner using q_i in the place of Q_i. If we now make the assumptions of Section 3, we may define $k = \frac{1}{2}(1+L)$ and $k_s = \frac{1}{2}(1+L_s)$. Then λ is given by the standard normal abscissa corresponding to k, regarded as an incomplete probability integral, and the Engel elasticity η_s of the specific commodity s is given by dividing the standard normal abscissa corresponding to k_s by λ. This method is slightly different from the one proposed elsewhere (Iyengar, 1960) and is not used in this paper although the former method appears to be intuitively better.

The several assumptions made in Section 3 have approximately been shown elsewhere to be empirically true. (See Iyengar 1960 and 1960a; and Roy and Dhar, 1959). In fact, it has been found from an analysis of data on consumer expenditure in rural and urban areas of India pertaining to the period (December 1955 —May 1956) that the expenditure distribution can well be described by a log-normal hypothesis with a relative standard deviation (λ) of 0.66 which gives a Lorenz ratio of 36 percent. This distribution is diagramatically represented in Figure 2 (Appendix) corresponding to $\beta = 0$. The changes in the distribution have been indicated by shifts in the Lorenz curve brought about by assumed changes in the level of inequality of incomes (total expenditures). Six different alternatives are considered as the likely increase in the average per capita expenditure and five distinct values are assumed for the changes in the concentration of income distribution. The constants associated with the expression for the percentage increase in demand have been set out in Table 2.

TABLE 2. VALUES OF $C(\beta,\eta)$ FOR DIFFERENT VALUES OF β AND η

item	elasticity η	proportionate reduction in concentration ratio (β)				
		−5%	5%	10%	15%	20%
(a)	0.58	0.9938	1.0053	1.0106	1.0154	1.0201
(b)	0.76	0.9954	1.0040	1.0079	1.0116	1.0150
(c)	1.55	1.0217	0.9816	0.9639	0.9477	0.9329
(d)	1.67	1.0286	0.9759	0.9528	0.9278	0.9128
(e)	1.95	1.0479	0.9604	0.9250	0.8898	0.8598

Using Table 2 we construct Table 3 giving percentage increase (or decrease) in demand for the five items under different sets of specifications regarding changes in the overall expenditure and concentration.

TABLE 3. PERCENTAGE INCREASE (OR DECREASE) IN DEMAND PER CAPITA

item	percentage increase in per capita total expenditure (100α)	percentage decrease in Lorenz ratio (100β)					
		−5	0	5	10	15	20
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(a) foodgrains	0	−0.6	0.0	0.5	1.1	1.5	2.0
	5	2.2	2.9	3.4	4.0	4.5	4.9
	10	5.0	5.7	6.2	6.8	7.3	7.8
	15	7.8	8.4	9.0	9.6	10.1	10.6
	20	10.5	11.2	11.7	12.3	12.9	13.4
	25	13.1	13.8	14.4	15.0	15.6	16.1
(b) all food	0	−0.5	0.0	0.4	0.8	1.2	1.5
	5	3.3	3.8	4.2	4.6	5.0	5.3
	10	7.0	7.5	7.9	8.4	8.8	9.1
	15	10.7	11.2	11.6	12.1	12.5	12.9
	20	14.3	14.9	15.3	15.8	16.2	16.6
	25	17.9	18.4	18.9	19.4	19.9	20.3
(c) clothing	0	2.2	0.0	−1.8	−3.6	−5.2	−6.7
	5	10.2	7.9	5.9	4.0	2.2	0.6
	10	18.4	15.9	13.8	11.7	9.9	8.1
	15	26.9	24.2	21.9	19.7	17.7	15.9
	20	35.5	32.7	30.2	27.9	25.7	23.8
	25	44.4	41.3	38.7	36.2	33.9	31.8
(d) milk and milk products	0	2.9	0.0	−2.4	−4.7	−7.2	−8.7
	5	11.6	8.5	5.9	3.4	0.7	−1.0
	10	20.6	17.3	14.4	11.7	8.8	7.0
	15	29.8	26.2	23.1	20.2	17.1	15.2
	20	39.5	35.6	32.2	29.2	25.8	23.8
	25	49.3	45.2	41.6	38.3	34.7	32.5
(e) medicine and medical services	0	4.8	0.0	−4.0	−7.5	−11.0	−14.0
	5	15.3	10.0	5.6	1.7	−2.1	−5.4
	10	26.2	20.4	15.7	11.4	7.2	3.5
	15	37.6	31.3	26.1	21.5	16.9	12.9
	20	49.5	42.7	37.1	32.0	27.0	22.7
	25	61.9	54.5	48.4	42.9	37.5	32.8

The interpretation of Table 3 is quite clear. For instance, assuming that the average income (total expenditure) per capita increases by 15 percent and the concentration of incomes becomes lower by 5 percent, the increase in demand per person for foodgrains, clothing and medicine are 9.0, 21.9 and 26.1 percent respectively.

If one did not consider the effect of changes in the Lorenz ratio one would get 8.4, 24.2 and 31.3 percent for the same items. In the latter case, the demand for clothing and medicine is to some extent exaggerated (by 3 and 7 percent respectively) while the demand for foodgrains is slightly underestimated (to the extent of 0.6 percent). It is shown in Section 3 that, in general, if we do not take into consideration the changes (towards Egalitarianism) in the distribution of incomes (total expenditures), the demand for necessary items would be underestimated and that for luxuries it would be somewhat overestimated.

If we do not consider the possibilities of a rise in the average level of total per capita expenditure, even then it is possible that the demand for necessaries augments under a more even distribution of incomes. For example, in the case of consumption of food the demand increases by 1.5 percent if the level of concentration is brought down by 20 percent (through taxation or appropriate policies of the Government). However, in this situation the demand for luxuries including clothing, milk and medical services, would shrink to a far greater extent than would the demand for necessaries increase. Medicine and medical services, being the most luxurious item among the ones we have considered in this paper, will suffer a loss of demand to the extent of 14 percent if the inequality of total expenditures were to diminish by 20 percent.

The following table illustrates the difference between estimates of increase in demand obtained by rising formulae (3.12) and (3.19) of Section 3. For the sake of simplicity, we shall assume $\beta = 0$. The estimates $I_{\alpha,0}\ (\eta)$ are given outside the brackets, the $I^*_{\alpha,0}\ (\eta)$ figures being inside.

TABLE 4. $I_{\alpha,0}\ (\eta)$ AND $I^*_{\alpha,0}\ (\eta)$ COMPARED

item	100α : percentage increase in overall expenditure per capita				
	5	10	15	20	25
(1)	(2)	(3)	(4)	(5)	(6)
(a) foodgrains	2.9 (2.9)	5.7 (5.8)	8.4 (8.7)	11.2 (11.6)	13.8 (14.5)
(b) food total	3.8 (3.8)	7.5 (7.6)	11.2 (11.4)	14.9 (15.2)	18.4 (19.0)
(c) clothing	7.9 (7.8)	15.9 (15.5)	24.2 (23.2)	32.7 (31.0)	41.3 (38.8)
(d) milk and milk products	8.5 (8.4)	17.3 (16.7)	26.2 (25.0)	35.6 (38.4)	45.2 (41.8)
(e) medicine and medical services	10.0 (9.8)	20.4 (19.5)	31.3 (19.5)	42.7 (39.0)	54.5 (48.8)

For necessary goods the difference between these estimates is not serious for lower values of α while the gap increases with the increase in the per capita income (total expenditure). A similar remark may be made regarding the relatively more luxurious items, with the difference that the gaps in these cases are wider.

For purposes of interpretation we shall produce a few interesting graphs in Figure 1 below corresponding to $\beta = 0$, $\beta = -0.05$ and $\beta = 0.10$ for just two items, one from the necessary group and another from the luxury group. These graphs show the percentage increase in demand per capita on the vertical axis and the percentage increase in income (total expenditure) per capita on the horizontal axis.

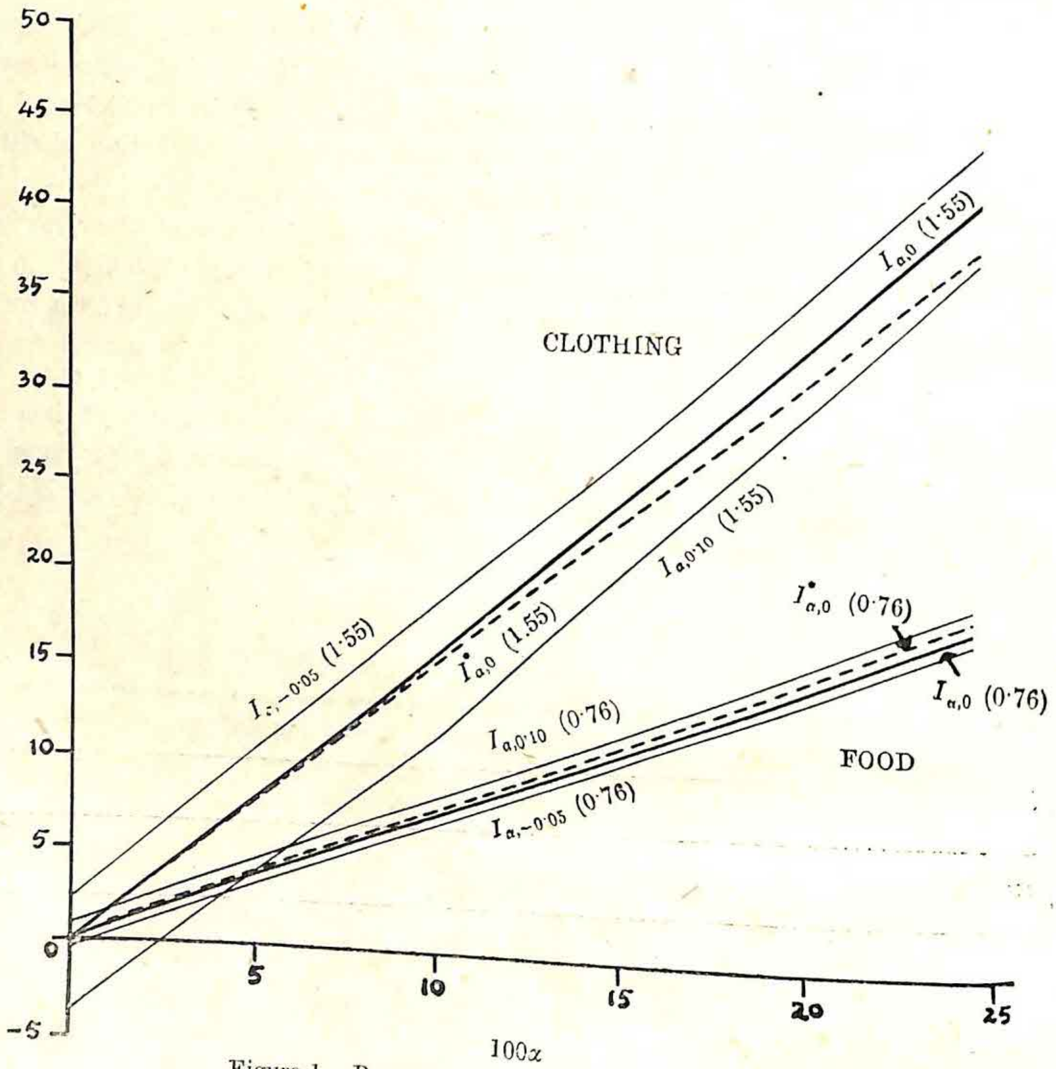


Figure 1. Percentage increase in demand per capita.

5. Acknowledgement

I am grateful to Dr. J. Roy, Shri M. Mukherjee and Shri N. Bhattacharya for their valuable suggestions and encouragement. I should also like to thank Professor Gerhard Tintner for useful comments.

Appendix

1. In Section 3 the initial value of L is taken to be 0.36 which, when substituted in the relation, $L = 2\phi\left(\frac{\lambda}{\sqrt{2}}\right) - 1$, leads to $\lambda = 0.66$. The new values of Lorenz ratios are computed by $L^* = (1-\beta)L$. From these the λ^*'s are computed for $\beta = -0.05$, $\beta = 0$ and $\beta = 0.10$ and are presented in Table A.1 below :

TABLE A.1. VALUES OF λ^* FOR DIFFERENT VALUES OF β

100β	L^*	λ^*
−5	0.378	0.6972
0	0.360	0.6600
10	0.324	0.5910

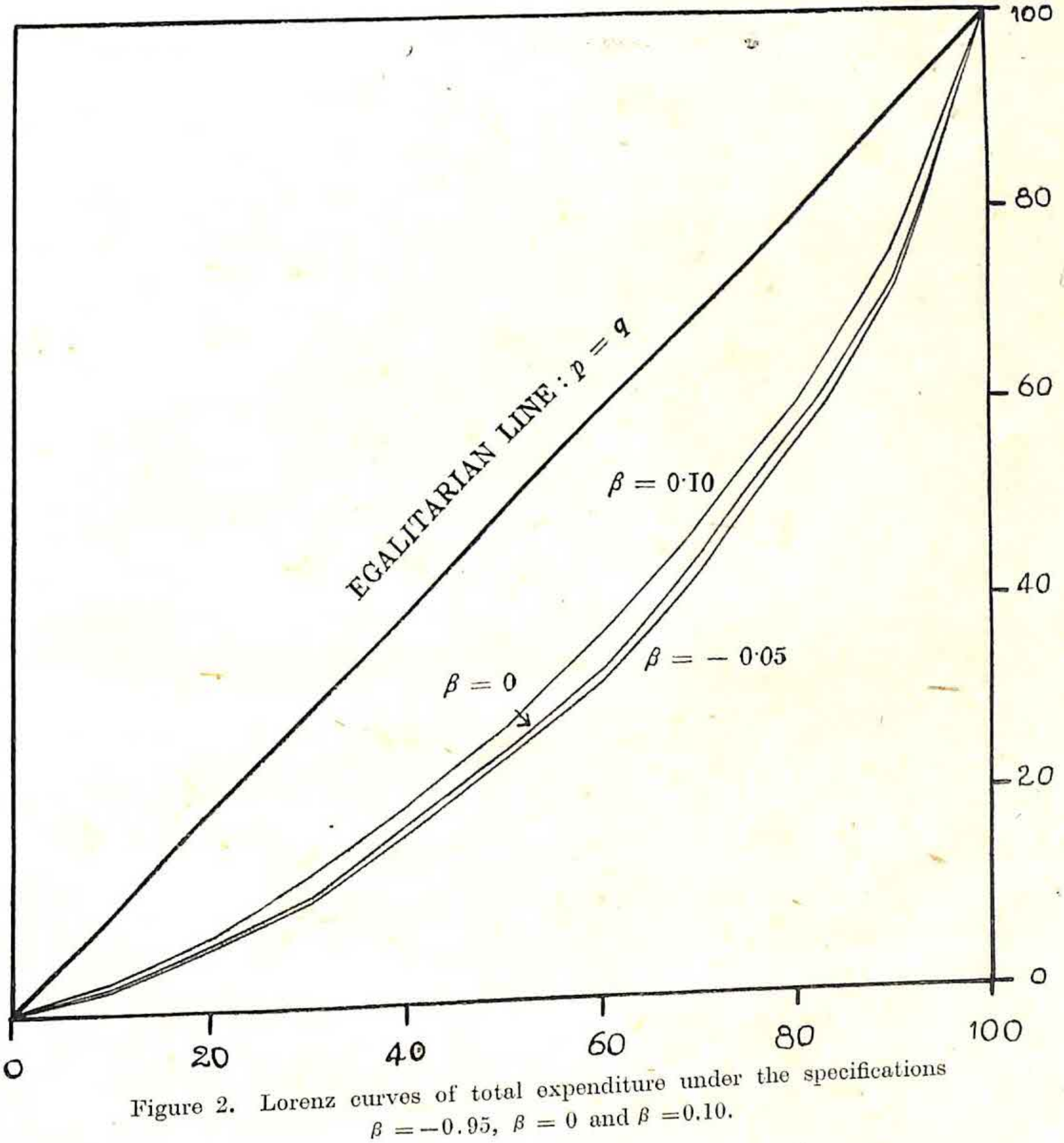


Figure 2. Lorenz curves of total expenditure under the specifications $\beta = -0.95$, $\beta = 0$ and $\beta = 0.10$.

2. Denoting by p the proportion of persons whose overall expenditure does not exceed a given amount, and by q the proportion of the total expenditure (of the community) spent by these persons, the Lorenz curve is drawn by taking p on the horizontal axis and q on the vertical axis. However, when the distribution of overall expenditures is known *a priori* to be log-normal the equation to the Lorenz curve may be written as

$$t_q = t_p - \lambda^* \qquad \ldots \quad (A.1)$$

where λ^* is the standard deviation of the logarithms of overall expenditures; t is, of course, defined in (3.9) of the text. The Lorenz curves of total expenditure are drawn in Figure 2 corresponding to the three distinct values of λ^* given in Table A.1.

REFERENCES

AITCHISON, J. and BROWN, J. A. C. (1957) : *The Log-normal Distribution with Special Reference to Its Uses in Economics*, Cambridge University Press.

ROY, J. and LAHA, R. G. (1959): Preliminary estimates of relative increase in consumer demand in rural and urban India. *Studies on Consumer Behaviour*, Asia Publishing House, Bombay and Statistical Publishing Society, Calcutta.

IYENGAR, N. S. (1960) : On a method of computing the Engel elasticities from concentration curves. (Accepted for publication in *Econometrica* **28**, No. 4. October 1960)

IYENGAR, N. S. (1960*a*) : On the standard error of the Lorenz concentration ratio. *Sankhyā*, **22**, 371-378.

ROY, J., CHAKRAVARTI, I. M., and LAHA, R. G. (1959) : A study of concentration curves as description of consumption pattern. *Studies on Consumer Behaviour*, Asia Publishing House, Bombay and Statistical Publishing Society, Calcutta.

ROY, J. and DHAR, S. K. (1959) : A study on the pattern of consumer expenditure in rural and urban India. *Studies on Consumer Behaviour*, Asia Publishing House, Bombay and Statistical Publishing Society, Calcutta.

Paper received : April, 1960.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Dr. R. R. Bahadur
Dr. D. Basu
Shri S. John
Dr. Sujit Kumar Mitra
Shrimati N. S. Nanjamma
Dr. C. R. Rao
Shri R. Ranga Rao
Dr. J. Roy
Shri J. Sethuraman

Editor,
Sankhyā : The Indian Journal of Statistics.

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

1. Place of Publication : 204/1 Barrackpore Trunk Road, Calcutta-35
2. Periodicity of its Publication : Quarterly
3. Printer's Name : Kalipada Mukherjee
   Nationality : Indian
   Address : 204/1 Barrackpore Trunk Road, Calcutta-35
4. Publisher's Name : Prafulla Chandra Mahalanobis
   Nationality : Indian
   Address : 204/1 Barrackpore Trunk Road, Calcutta-35
5. Editor's Name : P. C. Mahalanobis
   Nationality : Indian
   Address : 204 Barrackpore Trunk Road, Calcutta-35
6. Names and addresses of individuals who own the newspaper and partners or shareholders holding more than one per cent of the total capital : Published by Statistical Publishing Society (non-profit distributing agency)

I, Prafulla Chandra Mahalanobis, hereby declare that the particulars given above are true to the best of my knowledge and belief.

Date. 26 March, 1960

Sd/ Prafulla Chandra Mahalanobis

SANKHYĀ : THE INDIAN JOURNAL OF STATISTICS

CONTENTS OF VOLUME TWENTYTWO, 1960

(*All rights reserved*)

PAPERS

Page

Classification of natural and plantation teak (*Tectona Grandis*) grown at different localities of India and Burma with respect to its physical and mechanical properties. *By K. R. Nair and H. K. Mukherji* ... 1

A brief history of the organisation of official statistics in India during the British period. *By S. Subramanian* 85

Calculation of sampling errors for index numbers. *By K. S. Banerjee* ... 119

Type-study on peak period in harvesting *Aman* paddy, West Bengal, December 1954—January 1955. *By R. K. Som, G. C. Bhattacharyya and N. K. Namboodiri* 131

Next steps in planning. *By P. C. Mahalanobis* 143

Industrialization of underdeveloped countries—A means to peace. *By P. C. Mahalanobis* 173

Food resources and population of India—A historical study. *By P. C. Bansil* 183

Test scoring by the 602A calculating punch. *By Ajit Haldar* 209

The addition of random vectors. *By J. B. S. Haldane* 213

A note on the estimation of variance. *By Leo A. Goodman* 221

On the asymptotic efficiency of tests and estimates. *By R. R. Bahadur* ... 229

On bounded infinitely divisible random variables. *By Glen Baxter and J. M. Shapiro* 253

A limit theorem for densities. *By R. Ranga Rao and V. S. Varadarajan* ... 261

Tests for independence and symmetry in multivariate normal populations. *By J. Roy* 267

On the *F*-test in the intrablock analysis of a balanced incomplete block design. *By Sujit Kumar Mitra* 279

An admissible estimate for any sampling design. *By V. P. Godambe* ... 285

Non-null distribution of the likelihood-ratio in analysis of dispersion. *By J. Roy* 289

Nonparametric linear estimation of common median of symmetrical populations from symmetrically censored observations. *By John E. Walsh* 295

Page

On some classification problems–I. *By S. John* 301

On some classification statistics. *By S. John* 309

Multivariate analysis : An indispensable statistical aid in applied research. *By C. Radhakrishna Rao* 317

Maximum likelihood estimation for positively regular Markov chains. *By B. R. Bhat* 339

A note on the structure of a certain stochastic model. *By S. C. Das* ... 345

A short table of the hypergeometric function $_0F_1(c; x)$. *By J. Edward Jackson* 351

Approximate confidence interval for linear functions of means of K populations when the population variances are not equal. *By Saibal Kumar Banerjee* 357

Measurement of attrition in technical education. *By K. C. Seal* ... 359

A note on the efficiency of double sampling for stratification. *By M. V. Jambunathan* 367

On the standard error of the Lorenz concentration ratio. *By N. Sreenivasa Iyengar* 371

On a problem of estimating increase in consumer demand. *By N. Sreenivasa Iyengar* 379

REPORT

National Sample Survey : Number Thirteen : Report on Household Transport Operations 21

CORRIGENDA

Unbiassed minimum variance estimation in a class of discrete distributions. *By J. Roy and S. K. Mitra* 212

Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. *By M. N. Murthy and N. S. Nanjamma* 212

AUTHOR INDEX

Bahadur, R. R. On the asymptotic efficiency of tests and estimates ... 229

Banerjee, K. S. Calculation of sampling errors for index numbers ... 119

Banerjee, Saibal Kumar. Approximate confidence interval for linear functions of means of K populations when the population variances are not equal 357

Bansil, P. C. Food resources and population of India—A historical study 183

Baxter, Glen and Shapiro, J. M. On bounded infinitely divisible random variables 253

Page

Bhat, B. R. Maximum likelihood estimation for positively regular Markov chains ... 339

Bhattacharyya, G. C., Som, R. K. and Namboodiri, N. K. Type-study on peak period in harvesting *Aman* paddy, West Bengal, December 1954—January 1955 ... 131

Das, S. C. A note on the structure of a certain stochastic model ... 345

Goodman, Leo A. A note on the estimation of variance ... 221

Godambe, V. P. An admissible estimate for any sampling design ... 285

Haldane, J. B. S. The addition of random vectors ... 213

Haldar, Ajit. Test scoring by the 602A calculating punch ... 209

Iyengar, N. Sreenivasa. On the standard error of the Lorenz concentration ratio ... 371

——— On a problem of estimating increase in consumer damand ... 379

Jackson, J. Edward. A short table of the hypergeometric function ${}_0F_1(c; x)$ ... 351

Jambunathan, M. V. A note on the efficiency of double sampling for stratification ... 367

John, S. On some classification problems–I ... 301

——— On some classification statistics ... 309

Mahalanobis, P. C. Next steps in planning ... 143

——— Industrialization of underdeveloped countries—A means to peace 173

Mitra, Sujit Kumar. On the *F*-test in the intrablock analysis of a balanced incomplete block design ... 279

Mitra, S. K. and Roy, J. Corrigenda to a paper on "Unbiassed minimum variance estimation in a class of discrete distributions" ... 212

Mukherji, H. K. and Nair, K. R. Classification of natural and plantation teak (*Tectona Grandis*) grown at different localities of India and Burma with respect to its physical and mechanical properties ... 1

Murthy, M. N. and Nanjamma, N. S. Corrigenda to a paper on "Almost unbiased ratio estimates based on interpenetrating subsample estimates" ... 212

Nair, K. R. and Mukherji, H. K. Classification of natural and plantation teak (*Tectona Grandis*) grown at different localities of India and Burma with respect to its physical and mechanical properties ... 1

Namboodiri, N. K., Som, R. K. and Bhattacharyya, G. C. Type-study on peak period in harvesting *Aman* paddy, West Bengal, December 1954–January 1955 ... 131

Page

Nanjamma, N. S. and Murthy, M. N. Corrigenda to a paper on "Almost unbiased ratio estimates based on interpenetrating subsample estimates" 212

Rao, R. Ranga and Varadarajan, V. S. A limit theorem for densities ... 261

Rao, C. Radhakrishna. Multivariate analysis : An indispensable statistical aid in applied research 317

Roy, J. Tests for independence and symmetry in multivariate normal populations 267

——— Non-null distribution of the likelihood-ratio in analysis of dispersion 289

Roy, J. and Mitra, S. K. Corrigenda to a paper on "Unbiassed minimum variance estimation in a class of discrete distributions" ... 212

Seal, K. C. Measurement of attrition in technical education ... 359

Shapiro, J. M. and Baxter, Glen. On bounded infinitely divisible random variables 253

Some, R. K., Bhattacharyya, G. C. and Namboodiri, N. K. Type-study on peak period in harvesting *Aman* paddy, West Bengal, December 1954—January 1955. 131

Subramanian, S. A brief history of the organisation of official statistics in India during the British period 85

Varadarajan, V. S. and Rao, R. Ranga. A limit theorem for densities ... 261

Walsh, John E. Nonparametric linear estimation of common median of symmetrical populations from symmetrically censored observations 295